


Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions

Ewa Papajak, Jingjing Zheng, Xuefei Xu, Hannah R. Leverentz, and Donald G. Truhlar*

Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States

 Supporting Information

ABSTRACT: We present a perspective on the use of diffuse basis functions for electronic structure calculations by density functional theory and wave function theory. We especially emphasize minimally augmented basis sets and calendar basis sets. We base our conclusions on our previous experience with commonly computed quantities, such as bond energies, barrier heights, electron affinities, noncovalent (van der Waals and hydrogen bond) interaction energies, and ionization potentials, on Stephens et al.'s results for optical rotation and on our own new calculations (presented here) of polarizabilities and of potential energy curves of van der Waals complexes. We emphasize the benefits of partial augmentation of the higher-zeta basis sets in preference to full augmentation at a lower ζ level. Benefits and limitations of the use of fully, partially, and minimally augmented basis sets are reviewed for different electronic structure methods and molecular properties. We have found that minimal augmentation is almost always enough for density functional theory (DFT) when applied to ionization potentials, electron affinities, atomization energies, barrier heights, and hydrogen-bond energies. For electric dipole polarizabilities, we find that augmentation beyond minimal has an average effect of 8% at the polarized triple- ζ level and 5% at the polarized quadruple- ζ level. The effects are larger for potential energy curves of van der Waals complexes. The effects are also larger for wave function theory (WFT). Even for WFT though, full augmentation is not needed for most purposes, and a level of augmentation between minimal and full is optimal for most problems. The calendar basis sets named after the months provide a convergent sequence of partially augmented basis sets that can be used for such calculations. The jun-cc-pV(T+d)Z basis set is very useful for MP2-F12 calculations of barrier heights and hydrogen bond strengths.

INTRODUCTION

It has been realized for a long time^{1,2} that basis sets for electronic structure calculations of many molecular properties must contain diffuse basis functions for quantitative accuracy. It has also been shown many times that diffuse functions have a relatively small effect for a number of molecular properties (for example, for ionization potentials and stationary point geometries of many molecules). It is best to avoid using diffuse basis functions when they are unnecessary or to avoid using more than the required number when they are necessary because adding diffuse functions to a basis set not only increases the cost of the calculation but can cause problems with SCF convergence and can increase basis set superposition error (BSSE). However, systematic studies of how to include the diffuse space efficiently that are based on large sets of data for various molecular properties and barrier heights are few and far between.

We have recently conducted a series of such studies including barrier heights, electron affinities, ionization potentials, atomization energies, and hydrogen bond energies.^{3–6} We conclude that much current practice includes more diffuse functions than are needed. Often, better accuracy could be achieved if the additional cost were invested in higher- ζ basis set or more polarization functions. Our conclusions on how to achieve higher accuracy on a per-cost basis can be widely useful for practical electronic structure calculations, and so we summarize them in this article. In addition to considering our own results, we comment briefly on the optical rotation studies of Stephens, Sadlej, and co-workers.^{7–9}

In addition to reviewing and integrating previously published work, this paper presents new calculations of polarizability and potential energy curves, which provide interesting challenges for efficient basis set selection.

MINIMALLY AUGMENTED BASIS SETS

In the correlation consistent basis sets, Dunning and co-workers¹⁰ defined “aug-” (which denotes “augmented”) to mean adding one diffuse basis function to every atom for every angular momentum of basis functions already present on that atom. We call such basis sets fully augmented. For example, the cc-pVTZ basis set for methane has s, p, d, and f functions on C and s, p, and d functions on H, so aug-cc-pVTZ adds diffuse s, p, d, and f subshells to C and diffuse s, p, and d subshells to H. Dunning and co-workers individually optimized the exponential parameters for each aug- basis set.

In contrast, the earlier “plus” basis sets originally systematized by Pople and co-workers¹¹ added only diffuse s and p subshells to nonhydrogenic functions and no diffuse functions to hydrogen atoms. We call this minimal augmentation. In Pople-type basis sets, one uses the same exponential parameters for diffuse functions in any basis set for a given atom. For example, 6-31+G(d,p) and 6-311+G(2df,2p) use the same exponential parameters for diffuse functions on C.

We have defined three kinds of minimal augmentation, denoted by “+”, “maug-”, and “ma-”. We next explain these in turn.

Received: February 15, 2011

Published: August 05, 2011

The “+” kind of minimal augmentation can be applied to any basis set for any atom for which exponential parameters have been defined for adding diffuse functions to the 6-31G or 6-311G basis sets. We then add a plus suffix. For example, cc-pVTZ+ denotes a minimally augmented cc-pVTZ basis set with the exponential parameters for diffuse functions from ref 11. Some basis sets of the + type have other names. For example, the MG3S basis set¹² is minimally augmented, and in fact is the same as 6-311+G(3d2f,2df,2p) for H through Si but is improved¹³ for S, P, and Cl. The generalization of the + approach to basis sets other than 6-31+G and 6-311+G was introduced in ref 3.

The maug- kind of minimal augmentation can be applied to any aug- type correlation consistent basis set. One simply truncates the diffuse space to the minimal augmentation level. For example maug-cc-pVTZ retains the diffuse s and p functions on carbon with the exponential parameters optimized for the aug-case but deletes all other diffuse functions. This approach was introduced in refs 3 and 4. We note that other levels of augmentation between minimal (maug-) and full (aug-) have also been defined for correlation-consistent basis sets, and these are discussed in ref 6 and below.

The ma- kind of minimal augmentation⁵ may be used with any nondiffuse basis set. One simply creates diffuse functions for atoms heavier than He by dividing the smallest s and p exponential parameters already present by a factor of 3. The ma- type of minimal augmentation was originally proposed for the def2-basis sets,¹⁴ which are the second-generation default basis sets of the TURBOMOLE program, as developed in Karlsruhe. For example, the ma-TZVP basis set is obtained from the def2-TZVP basis set for carbon by adding diffuse functions obtained this way.⁵ For basis sets other than the Karlsruhe def2- series, one simply adds ma- as a prefix, e.g., ma-LANL2DZ.

Examples of applications of the efficient “+” and “maug-” basis sets to problems involving large molecules are available both in work by our group^{15–17} and in work by other research groups.^{18,19}

It is sometimes assumed that adding extra basis functions is a useful cautionary step, so it does not hurt to use aug- when a smaller basis set would be sufficient. However, using full augmentation raises the cost, and if it is done at the expense of not increasing the size of the valence space, it can be harmful. For example, it is often more beneficial to increase a basis set from maug-cc-pVDZ to maug-cc-pVTZ rather than increase it to aug-cc-pVDZ, even for properties that are sensitive to including diffuse functions, for example, barrier heights, hydrogen bond energies, and electron affinities. When applying density functional theory, the only exceptions to this general finding in our tests were B3LYP calculations of hydrogen bonding energies, which also showed no improvement in accuracy from augmentation of the quadruple- ζ basis set, and which showed higher accuracy of maug than aug at the triple- ζ level; these findings indicate that the error is dominated for B3LYP by the intrinsic error in the density functional, not by the lack of completeness of the basis set.

There are also other ways that adding extra diffuse basis functions could be harmful. For example, it can increase basis set superposition error. It is sometimes speculated that adding more functions to the basis set increases its completeness and should therefore reduce the basis set superposition error; however, it is especially the functions that overlap other centers that cause BSSE, so it is also possible to increase BSSE by adding diffuse functions. We have studied this and have shown³ that the increased diffuse space of the fully augmented basis sets does not

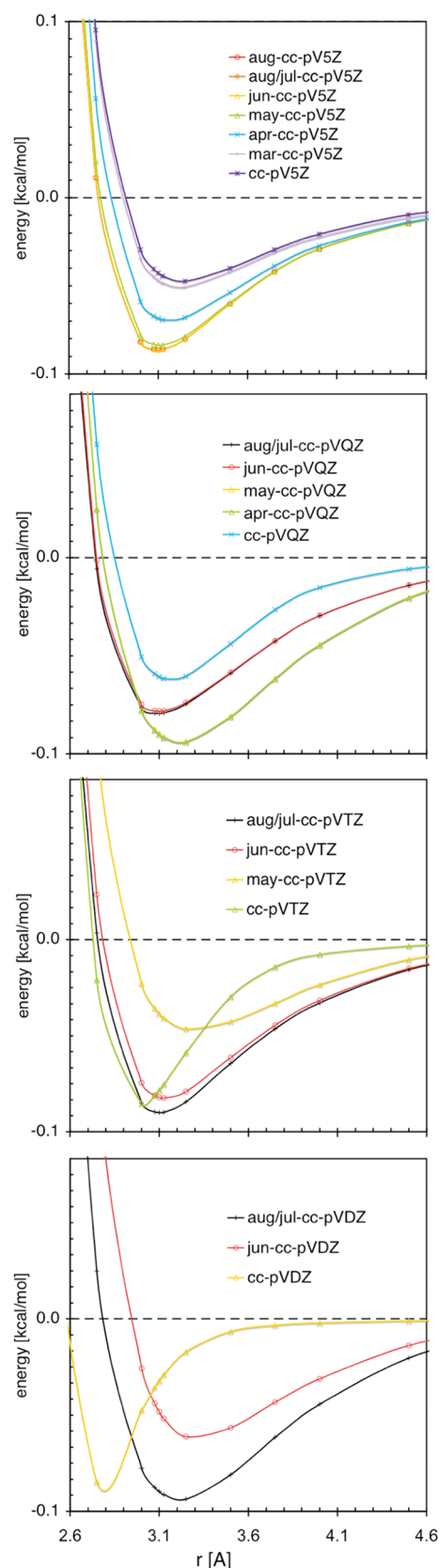


Figure 1. CCSD(T) interaction energy (in kcal/mol) for Ne_2 relative to the energy of Ne atoms at infinite separation for quintuple-, quadruple-, triple-, and double- ζ basis sets. The abscissa r is the distance between the Ne atoms. No counterpoise corrections are applied in Figures 1–6.

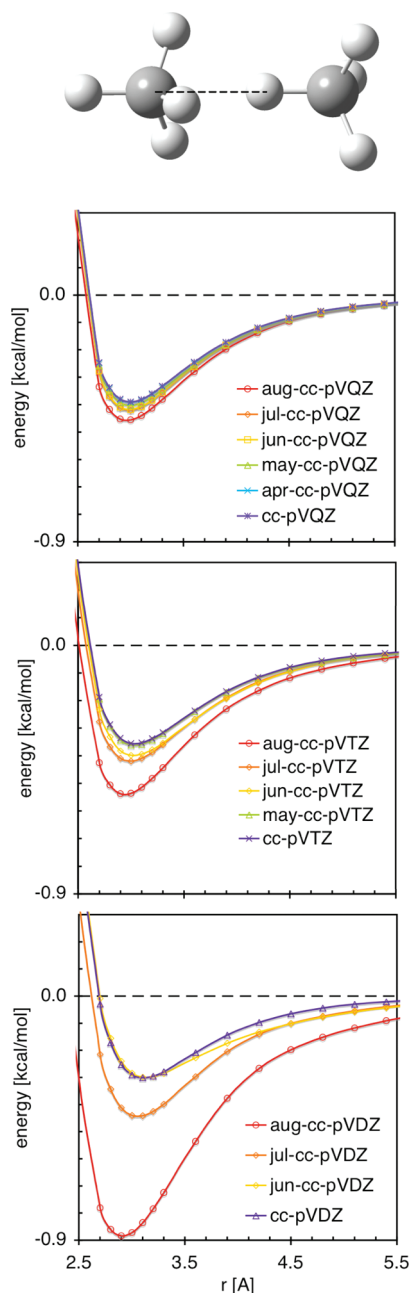


Figure 2. CCSD(T) interaction energy (in kcal/mol) for the CH_4 dimer relative to the energy of two CH_4 molecules at infinite separation for quadruple-, triple-, and double- ζ basis sets. The abscissa r is the distance between the C and H atoms, as indicated by a dashed line (---) in the figure. The structures have C_{3v} symmetry.

necessarily decrease the counterpoise correction for basis set superposition error, and in fact it can increase it. One could hypothesize that this is an indication of the limitations of counterpoise corrections, not necessarily increased BSSE. In any case, one can conclude from this that an expensive full augmentation followed by an also expensive counterpoise correction calculation is not necessarily accurate, besides being uneconomical.

DENSITY FUNCTIONAL THEORY

We have found^{3–5} that minimal augmentation is almost always enough for density functional theory (DFT) when applied to

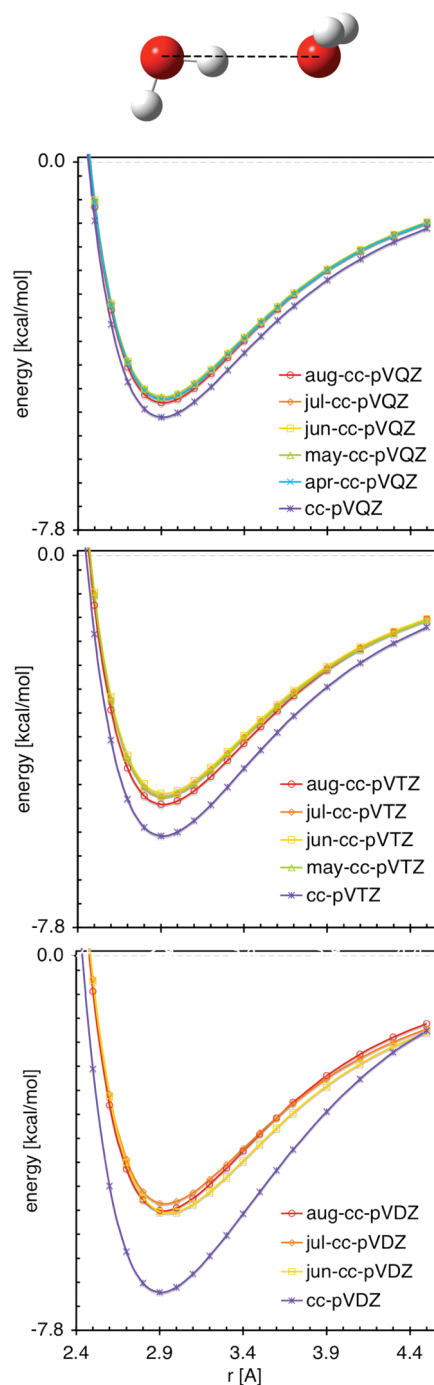


Figure 3. CCSD(T) interaction energy (in kcal/mol) for $(\text{H}_2\text{O})_2$ relative to the energy of H_2O molecules at infinite separation for quadruple-, triple-, and double- ζ basis sets. The abscissa r is the distance between the O atoms as indicated by a dashed line (---) in the figure.

ionization potentials, electron affinities, atomization energies, barrier heights, and hydrogen-bond energies. Indeed, it has advantages over full augmentation not only in cost but also in sometimes reducing slow SCF convergence and basis set superposition error. We also note that minimal augmentation (relative to no augmentation) is often more advantageous in DFT than in wave function theory (WFT).²⁰ Minimal augmentation is especially important for barrier heights, hydrogen bonding, and electron affinities but less important for bond energies and ionization potentials.³

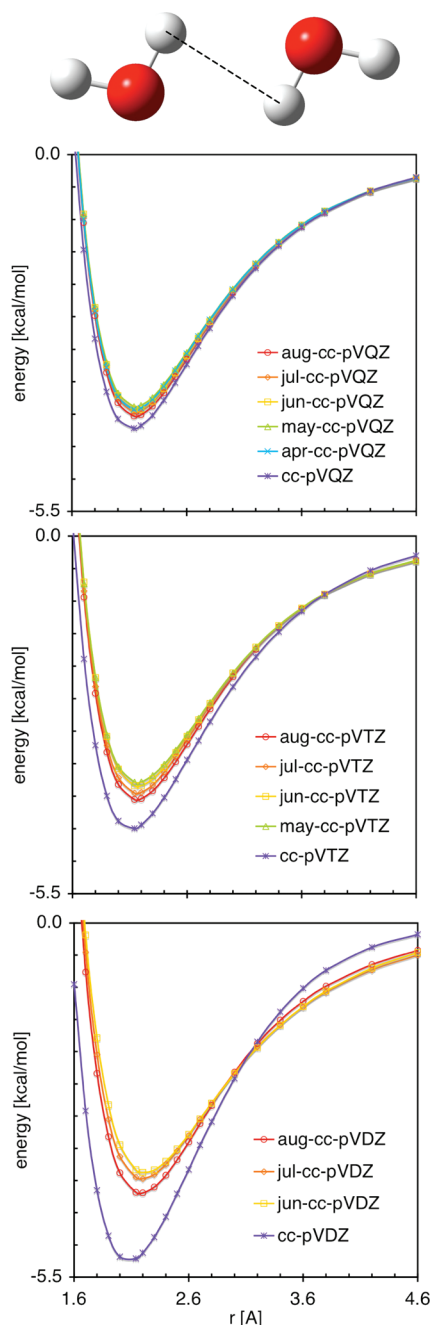


Figure 4. CCSD(T) interaction energy (in kcal/mol) for $(\text{H}_2\text{O})_2$ relative to the energy of $(\text{H}_2\text{O})_2$ molecules at infinite separation. The abscissa r is the distance between two H atoms as indicated by a dashed line (---) in the figure.

In DFT calculations, the savings from using minimal augmentation instead of full augmentation can be significant since DFT is often applied to challenging problems such as large systems or direct dynamics calculations that require a large number of electronic structure calculations on the fly. In such cases, it is very undesirable to increase the cost of the calculations by using more than the minimum number of basis functions required for chemical accuracy. Raising the cost by overaugmentation may make it necessary to adopt compromises such as a reduced number of steps in dynamics calculations, using a smaller than necessary grid, using a lower- ζ basis set, etc. The other extreme is

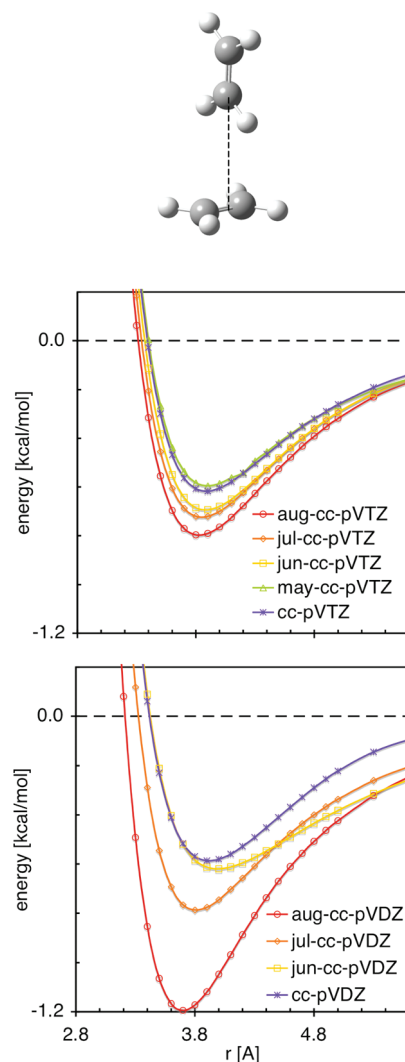


Figure 5. CCSD(T) interaction energy (in kcal/mol) for $(\text{C}_2\text{H}_4)_2$ relative to the energy of C_2H_4 molecules at infinite separation. The abscissa r is the distance indicated by a dashed line (---) in the figure.

the omission of all diffuse functions because of the cost. *Our default recommendation for DFT is to use minimally augmented basis sets instead of full augmentation for properties such as barrier heights, electron affinities, polarizabilities, and noncovalent interactions; however, no augmentation is needed for ionization potentials and atomization energies.*

WAVE FUNCTION THEORY AND CALENDAR BASIS SETS

Correlated wave function calculations are more sensitive to diffuse basis set expansion beyond the minimum than are density functional calculations, and therefore we recommend higher than minimal expansions of the diffuse space but still recommend less than full augmentation in most cases.

The reason for the slower convergence of correlated wave function calculations, when compared to density functional calculations, is easily understood. When including dynamic electron correlation in WFT, one uses the basis functions not only to represent the single-particle density but also to build the cusps into the many-electron wave function in the region where two

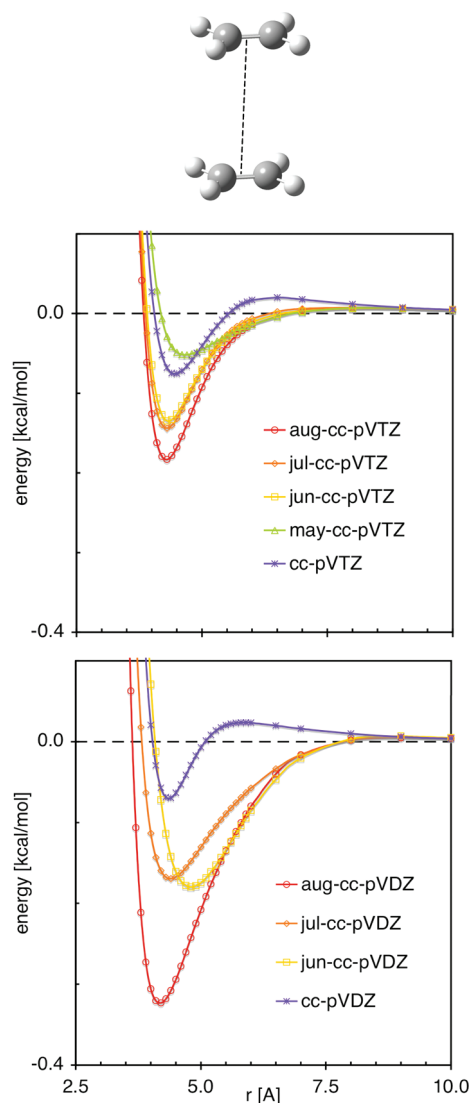


Figure 6. CCSD(T) interaction energy (in kcal/mol) for $(\text{C}_2\text{H}_4)_2$ relative to the energy of C_2H_4 molecules at infinite separation. The abscissa r is the distance indicated by a dashed line (---) in the figure.

electrons come together. The cusp is built up from a large number of configurations, extending to fairly high angular momentum for individual orbitals but with each configuration having a small coefficient, and hence second-order perturbation theory (MP2) already provides realistic predictions of these coefficients^{21,22} (the higher-order correlations are important but are less dominant in the description of the cusps). As a consequence, the basis set requirements may be examined at the MP2 level. We therefore recently studied the effect of diffuse functions on basis set convergence of MP2 calculations, and we developed a convergent sequence of partially augmented convergent correlation-consistent basis sets named after the months.⁶

In these studies, we found that even in WFT, just as in DFT, for all practical purposes one does not need diffuse functions on hydrogen atoms. Thus, our first recommendation is to delete these on all hydrogen and helium atoms; this yields basis sets we label as jul-, a naming convention that we actually introduced in a previous paper.³ (This is sometimes called 'aug', but 'aug' has been used to denote more than one possible

modification to the aug- basis sets, so we believe that a unique notation is useful.)

The other calendar basis sets are obtained by successively deleting diffuse subshells on heavier atoms.⁶ Consider carbon. The aug- and jul- triple- ζ basis sets have diffuse s, p, d, and f subshells. We delete the f subshell to get jun-, then the d subshell to get may-, which is equivalent to maug- for triple- ζ . But at the quadruple- ζ level, carbon has diffuse s, p, d, f, and g diffuse subshells. We delete the g subshell to get jun-, the f subshell to get may-, and the d subset to get apr-, which is equivalent to maug- for quadruple- ζ . Thus, when one increases the basis from jun-cc-pVTZ to jun-cc-pVQZ, one not only increases the valence basis but also adds a diffuse f subshell; in this sense, the diffuse space is convergent when one increases the ζ level but keeps the month constant (unlike the case in going from 6-31+G(d,p) to 6-311+G(2df,2p)). Thus, for some problems, the calendar basis sets can provide a more efficient sequence of basis sets (than unaugmented, minimally augmented, or fully augmented sets) for basis set extrapolation to the complete basis set limit.

The lowest exponents in the unaugmented cc-pV x Z basis sets decrease as x increases. Therefore, even though unaugmented, the underlying cc-pV x Z basis sets become more and more diffuse. This is one of the reasons why—for the previously published data on barrier heights, electron affinities, and hydrogen bonding interactions—the need for the diffuse functions decreases as the valence space increases. For DZ, we recommended jul-, while for QZ we can reduce the basis set to jun- or even may- without a significant loss in accuracy.

In general, the goal of most electronic structure calculations is not to achieve an accuracy on the order of, for example, 0.1 kcal/mol at any cost but rather to achieve as high an accuracy possible at a reasonable cost. The actual accuracy objective is dictated by the system's size and character and by the electronic structure method itself. For example, it is unreasonable to invest an additional $\sim 30\%$ of computational time to improve accuracy by about 0.1 kcal/mol when the expected accuracy of the method is ~ 1 kcal/mol and the additional cost would have been better invested in increasing other features of the basis set or using a different level of theory. The need for this kind of approach is evidenced by the enormous number of calculations in the literature that employ double- ζ and triple- ζ basis sets, even though these calculations are clearly not at the basis set limit.

We recommend jun- basis sets as a better default option than aug- basis sets for correlated WFT calculations. On the basis of our tests, barrier heights, hydrogen bonding, electron affinities, ionization potentials, and atomization energies,⁶ it is generally more beneficial, on a performance versus cost basis, to invest additional computational resources in increasing a basis set from cc-pVDZ to jun-cc-pVTZ rather than to aug-cc-pVDZ or cc-pVTZ. The analogous statement is true for the jun-cc-pVTZ to may-cc-pVQZ transition (skipping the aug-cc-pVTZ and cc-pVQZ). We note that, averaged over a large database, aug- triple- ζ calculations involve about 27% more basis functions than jun- triple- ζ .⁶

An alternative way to converge correlated WFT calculations with respect to the basis set is the F12 method,^{24,25} in which basis functions depending explicitly on interelectronic distances are added to the basis. This is a powerful method because it eliminates much of the need to build up the electron–electron cusps by a succession of higher-angular-momentum basis functions. This method is expected to revolutionize quantum chemistry over the next several years by allowing nearly complete-basis-set calculations with much less effort than was previously required.

Table 1. HF and MP2 Polarizability [A^3] Values, Average Unsigned Errors [A^3], and Percentage Errors [%] for Basis Sets of Different Degrees of Augmentation^a

	H ₂ O		CH ₄		C ₅ H ₅ N		HO(CH ₂) ₃ SH		mean unsigned error		mean unsigned percentage error	
	HF	MP2	HF	MP2	HF	MP2	HF	MP2	HF	MP2	HF	MP2
aug-cc-pV(Q+d)Z	1.26	1.43	2.37	2.45	9.43	9.61	9.27	9.71	0.00	0.00	0.0	0.0
jul-cc-pV(Q+d)Z	1.24	1.40	2.36	2.43	9.43	9.61	9.26	9.74	0.01	0.02	0.4	0.7
jun-cc-pV(Q+d)Z	1.24	1.40	2.36	2.43	9.43	9.61	9.26	9.73	0.01	0.02	0.4	0.7
may-cc-pV(Q+d)Z	1.24	1.39	2.36	2.43	9.43	9.59	9.26	9.71	0.01	0.02	0.5	0.9
apr-cc-pV(Q+d)Z ^b	1.16	1.28	2.27	2.32	9.31	9.46	9.10	9.52	0.12	0.15	3.8	4.7
cc-pV(Q+d)Z	1.09	1.17	2.27	2.31	9.01	9.13	8.85	9.19	0.28	0.35	6.7	8.4
aug-cc-pV(T+d)Z	1.25	1.41	2.37	2.45	9.43	9.63	9.25	9.76	0.01	0.02	0.3	0.5
jul-cc-pV(T+d)Z	1.20	1.33	2.35	2.42	9.42	9.61	9.22	9.69	0.04	0.03	1.6	1.9
jun-cc-pV(T+d)Z	1.20	1.33	2.35	2.41	9.42	9.60	9.21	9.67	0.04	0.05	1.6	2.2
may-cc-pV(T+d)Z ^b	1.09	1.19	2.19	2.23	9.19	9.34	8.88	9.26	0.24	0.29	7.0	8.2
cc-pV(T+d)Z	0.98	1.03	2.18	2.20	8.59	8.63	8.41	8.66	0.54	0.65	12.2	14.6
aug-cc-pV(D+d)Z	1.20	1.36	2.36	2.44	9.40	9.63	9.17	9.67	0.05	0.03	1.5	1.4
jul-cc-pV(D+d)Z	1.09	1.20	2.28	2.33	9.36	9.57	9.02	9.44	0.15	0.17	5.2	6.1
jun-cc-pV(D+d)Z ^b	0.92	1.02	1.99	2.01	8.91	9.08	8.36	8.67	0.54	0.60	14.6	15.7
cc-pV(D+d)Z	0.74	0.76	1.91	1.91	7.77	7.83	7.53	7.63	1.09	1.27	24.2	27.2

^aThe methane (CH₄) and water (H₂O) geometries were optimized at the QCISD/MG3S level, while 3-mercaptoopropan-1-ol (HO(CH₂)₃SH) and pyridine (C₅H₅N) geometries were optimized at the MP2-F12/aug-cc-pV(Q+d)Z level with frozen core electrons. Note that for elements up through Mg, cc-pV(X+d)Z is the same as cc-pVXZ. ^bSame as maug-.

For F12 calculations, we recommend may- quadruple- ζ , jun- triple- ζ , and jul- double- ζ .⁶

The next section considers wave function calculations on properties not included in our previous systematic tests, and we will examine whether some of these properties are in greater need of diffuse functions than the properties considered so far.

■ WAVE FUNCTION CALCULATIONS OF VAN DER WAALS POTENTIAL CURVES, POLARIZABILITIES, RAMAN INTENSITIES, AND OPTICAL ROTATION

Accurate calculations of electrical properties of molecules often require large basis sets, including multiple polarization functions and diffuse functions.^{26–36} It is therefore interesting to study the need for higher-angular-momentum diffuse functions, and we present some calculations exploring this need in this section.

The long-range portions of the van der Waals potential curves between nonpolar molecules depend on induced low-order electric multipole moments, and therefore they provide a challenging test of the need for diffuse basis functions. Here we report new calculations designed to examine this issue. Potential energy curves were calculated for Ne₂, (CH₄)₂, (H₂O)₂, and (C₂H₄)₂ by using coupled cluster theory with single and double excitations and a quasiperturbative treatment of connected triple excitations³⁷ [CCSD(T)] using the *Gaussian 09*³⁸ program package; the results are presented in tabular form in the Supporting Information and are plotted in Figures 1–6. The geometries of the methane dimer, two orientations of the water dimer (nonplanar C_s and cyclic C_{2h}), and two orientations of the ethylene dimer (T-shaped and stacked) are shown in the figures. The potential energy curves show that diffuse functions play a larger role for van der Waals potentials than they do for the properties examined earlier in this perspective. Nevertheless, some savings are possible. For example, jun-QZ and jun-TZ agree reasonably well with aug-QZ and aug-TZ, respectively.

Table 2. HF, MP2, and MP2-F12 Finite-Field Approximation to zz Element of the Polarizability Tensor [A^3] of H₂O for Basis Sets of Different Degrees of Augmentation

	H ₂ O		
	HF	MP2	MP2-F12
jul-cc-pV(5+d)Z	1.25	1.40	1.41
aug-cc-pV(Q+d)Z	1.26	1.42	1.42
jul-cc-pV(Q+d)Z	1.24	1.38	1.38
jun-cc-pV(Q+d)Z	1.24	1.38	1.38
may-cc-pV(Q+d)Z	1.24	1.37	1.38
apr-cc-pV(Q+d)Z	1.14	1.24	1.29
cc-pV(Q+d)Z	1.11	1.19	1.25
aug-cc-pV(T+d)Z	1.24	1.40	1.41
jul-cc-pV(T+d)Z	1.19	1.31	1.33
jun-cc-pV(T+d)Z	1.18	1.30	1.33
may-cc-pV(T+d)Z	1.06	1.15	1.23
cc-pV(T+d)Z	1.00	1.05	1.16
aug-cc-pV(D+d)Z	1.19	1.34	1.38
jul-cc-pV(D+d)Z	1.06	1.15	1.22
jun-cc-pV(D+d)Z	0.88	0.95	1.13
cc-pV(D+d)Z			

Figure 3 shows particularly rapid convergence with respect to increasing the number of diffuse functions, with all curves from maug- to aug- being closely grouped at all three ζ levels, although they are far from the unaugmented curves. The curves in Figure 4 also converge rapidly.

Another problem for which full augmentation is sometimes recommended is the calculation of electric dipole polarizabilities. First, we used the analytic polarizability algorithm in *MolPro 2009*³⁹ to calculate HF and MP2 spherically averaged polarizabilities

Table 3. HF, MP2, and MP2-F12 Finite-Field Approximation to zz Element of the Polarizability Tensor [\AA^3] of CH_4 for Basis Sets of Different Degrees of Augmentation

	CH_4		
	HF	MP2	MP2-F12
aug-cc-pV(T+d)Z	2.37	2.45	2.44
jul-cc-pV(T+d)Z	2.35	2.42	2.42
jun-cc-pV(T+d)Z	2.35	2.42	2.42
may-cc-pV(T+d)Z	2.19	2.23	2.33
cc-pV(T+d)Z	2.18	2.21	2.31
aug-cc-pV(D+d)Z	2.36	2.44	2.44
jul-cc-pV(D+d)Z	2.28	2.33	2.37
jun-cc-pV(D+d)Z	1.99	2.01	2.24
cc-pV(D+d)Z	1.91	1.91	2.18

for four molecules: methane (CH_4), water (H_2O), 3-mercapto-propan-1-ol ($\text{HO}(\text{CH}_2)_3\text{SH}$), and pyridine ($\text{C}_5\text{H}_5\text{N}$). These results are in Table 1. Then, we used the finite field approximation with a field strength of 0.005 atomic units to calculate the zz element of the polarizability tensor for water and methane by the HF, MP2, and MP2-F12 methods. These results are in Tables 2 and 3.

Table 1 shows that diffuse functions on the H atoms have a significant effect on the results (more than a couple of percent) in the case of double- ζ basis sets for MP2. However, for TZ and QZ, the average difference between aug- and jul- is only 0.04 and 0.02 \AA^3 , respectively. For nonhydrogenic atoms, as in the case of properties considered in earlier sections of this perspective, we find that, at the lower- ζ levels, the higher angular momentum functions affect the results significantly. However, the lower angular momentum functions in jun-TZ and jul-DZ provide most of the difference between augmented and nonaugmented basis sets. At higher ζ levels, full augmentation only adds additional cost and does not affect the accuracy to any significant degree. For example, the results obtained with the may-QZ basis set agree well with those obtained with aug-QZ with the largest percentage error being 2% for H_2O .

Tables 2 and 3 allow us to compare MP2-F12 convergence to MP2 convergence for polarizability elements. For the MP2-F12 calculations of electric dipole polarizabilities, one finds that may-, jun-, and jul- quadruple- ζ basis sets all have mean unsigned percentage errors (with respect to the largest fully augmented basis set) of about 2.7% and less; jul- and jun- triple- ζ basis sets have mean unsigned percentage errors less than 5.7%; and jul- double- ζ has a mean unsigned percentage error (with respect to aug-) of 11%.

We conclude that diffuse functions are especially important in polarizability calculations employing double- ζ basis sets, but partially augmented basis sets of higher- ζ levels can be useful for polarizabilities. This is consistent with findings of Sadlej et al. who introduced one diffuse function for each shell in their single contracted basis sets for electric properties, while their large basis sets are augmented only with a single s on hydrogenic atoms and a single s and a single p on other elements.^{40,41} These basis sets have been found to perform well for Raman intensities by Schlegel,⁴² which is also consistent with our results since Raman intensities depend on the derivatives of polarizabilities with respect to nuclear coordinates. However, Zuber and Hug⁴³ find that full augmentation on heavy atoms and additional p functions on the hydrogen atoms are necessary for the accurate

description of Raman and Raman optical activity (ROA) scattering.

Optical rotation is an example of a property, like van der Waals potential curves, that is especially sensitive to the diffuse space of the basis set. Diffuse functions are necessary for good results; however, the most thorough studies suggest that some savings can be made. For example, in refs 7 and 9, the authors state that their HF and DFT 6-311++G(2d,2p) results, which are augmented but much less than fully augmented, agree well with results obtained by aug-cc-pVTZ.

SUMMARY

Diffuse functions are an important component of basis sets, but in most cases, we recommend less than full augmentation with diffuse functions. For density functional theory, we usually recommend minimal augmentation. The need for diffuse functions is both greater and more variable in correlated wave function theory, so it is harder to make a general recommendation, but a good general starting point is a jun- level of augmentation.

ASSOCIATED CONTENT

S Supporting Information. Tables of the potential energy curves for van der Waals complexes. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: truhlar@umn.edu.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under grant no. CHE09-56776.

REFERENCES

- Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 264.
- Davidson, E. R.; Feller, D. *Chem. Rev.* **1986**, *86*, 681.
- Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1197. **2009**, *5*, 3330(E).
- Papajak, E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 597.
- Zheng, J.; Xu, X.; Truhlar, D. G. *Theor. Chem. Acc.* **2011**, *128*, 295.
- Papajak, E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 10.
- Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J. *J. Phys. Chem. A* **2001**, *105*, 5356.
- Baranowska, A.; Laczowski, K. Z.; Sadlej, A. J. *J. Comput. Chem.* **2010**, *31*, 1176.
- Cheeseman, J. R.; Frisch, M. J.; Devlin, F. J.; Stephens, P. J. *J. Phys. Chem. A* **2000**, *104*, 1039.
- Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6769.
- Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.
- Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett.* **1999**, *306*, 407.
- Curtiss, L. A.; Raghavachari, K.; Redfern, C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
- F. Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- Zheng, J.; Truhlar, D. G. *J. Phys. Chem. A* **2009**, *113*, 11919.

- (16) Zheng, J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7782.
- (17) Yang, K.; Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2010**, *132*, 164117.
- (18) Meyer, M. M.; Kass, S. R. *J. Org. Chem.* **2010**, *75*, 4274.
- (19) Hermosilla, L.; Catak, S.; Van Speybroeck, V.; Waroquier, M.; Vandenberghe, J.; Motmans, F.; Adriaensens, P.; Lutsen, L.; Cleij, T.; Vanderzande, D. *Macromolecules* **2010**, *43*, 7424.
- (20) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (21) Petersson, G. A.; Braunschweig, M. *J. Chem. Phys.* **1985**, *83*, 5129.
- (22) East, A. L. L.; Allen, W. D. *J. Chem. Phys.* **1993**, *99*, 4638.
- (23) Sullivan, M. B.; Iron, M. A.; Redfern, P. C.; Martin, J. M. L.; Curtiss, L. A.; Radom, L. *J. Phys. Chem. A* **2003**, *107*, 5617.
- (24) Ten-no, S. *Chem. Phys. Lett.* **2004**, *398*, 56.
- (25) Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.
- (26) Werner, H. J.; Meyer, W. *Mol. Phys.* **1976**, *31*, 855.
- (27) Werner, H. J.; Meyer, W. *Phys. Rev. A* **1976**, *13*, 13.
- (28) Morrison, M. A.; Hay, P. J. *J. Phys. B* **1977**, *10*, L647.
- (29) Eades, R. A.; Truhlar, D. G.; Dixon, D. A. *Phys. Rev. A* **1979**, *20*, 867.
- (30) Douglass, C. H., Jr.; Weil, D. A.; Eades, R. A.; Truhlar, D. G.; Dixon, D. A. In *Chemical Applications of Atomic and Molecular Electrostatic Potentials*; Politzer, P., Truhlar, D. G., Eds.; Plenum: New York, 1981; p 173.
- (31) Darling, C. L.; Schlegel, H. B. *J. Phys. Chem.* **1994**, *98*, 5855.
- (32) Peterson, K. A.; Dunning, T. H., Jr. *THEOCHEM* **1997**, *400*, 93.
- (33) Cybulski, A. M. *J. Chem. Phys.* **1999**, *111*, 10520.
- (34) Oliveira, L. N.; Amaral, O. A. V.; Castro, M. A.; Fonseca, T. L. *Chem. Phys.* **2003**, *289*, 221.
- (35) Arruda, P. M.; Neto, A. C.; Jorge, F. E. *Int. J. Quantum Chem.* **2009**, *109*, 1189.
- (36) Camiletti, G. G.; Canal Neto, A.; Jorge, F. E.; Machado, S. F. *THEOCHEM* **2009**, *910*, 122.
- (37) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.
- (39) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrennar, T.; Knizia, G.; Koppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2009.1; University College Consultants Ltd.: Cardiff, Wales, U.K., 2009.
- (40) Benkova, Z.; Sadlej, A. J.; Oakes, R. E.; Bell, S. E. *J. Comput. Chem.* **2005**, *26*, 145.
- (41) Baranowska, A.; Sadlej, A. J. *J. Comput. Chem.* **2010**, *31*, 552.
- (42) Halls, M. D.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 8819.
- (43) Zuber, G.; Hug, W. *J. Phys. Chem.* **2004**, *108*, 2108.

Singular Value Decomposition for Analyzing Temperature- and Pressure-Dependent Radial Distribution Functions: Decomposition into Grund RDFs (GRDFs)

Philipp J. di Dio,* Martin Brehm, and Barbara Kirchner

Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Universität Leipzig, Linnéstr. 2, D-04103 Leipzig, Germany

ABSTRACT: Singular value decomposition paves the way for systematic investigations of temperature- and pressure-dependent radial distribution functions. The decomposition into (weighted) Grund radial distribution functions (GRDF) shows that the temperature-dependent water structure can easily be understood by only three contributions: a major temperature-independent contribution from the first GRDF, a major temperature-dependent contribution from the second GRDF, and a minor temperature-dependent fine structure contribution from the third GRDF.

We present a new and unbiased way of analyzing temperature as well as pressure- or density-dependent radial distribution functions (RDFs). For many decades, radial distribution functions of liquids, especially water, were investigated experimentally as well as theoretically to gain information about the structure of liquids and, in special cases, their temperature as well as pressure dependencies.^{1,2} For instance, the two-state model of water² is based on such investigations. Unfortunately, up to now, the analyses of the temperature- and pressure-dependent RDFs were restricted to the changing forms of the RDFs and the actual measured phase point.

Our approach of analyzing temperature- and pressure-dependent radial distribution functions is to use the singular value decomposition (SVD) of a matrix. The matrix \mathbf{A} is formed by the temperature- and/or pressure-dependent RDFs (columns a_i of \mathbf{A}) and then is decomposed with the SVD to determine the linear dependency of the RDFs. We found that five O–O RDFs (obtained from ab initio molecular dynamics simulations of water at 300, 350, 400, 600, and 1000 K) are, in a first approximation, a linear combination of only two (main) contributions:

$$g_{\text{O-O}}(r, T) \approx v_1(T) \cdot w_1(r) + v_2(T) \cdot w_2(r) \quad (1)$$

The analysis of our set of RDFs shows that $v_1(T)$ is almost constant, and the whole temperature dependency is condensed in $v_2(T) \cdot w_2(r)$. A possible application could lie in coarse-grained force fields, which often suffer from a lack of temperature transferability of the iterative Boltzmann inversion.³ The iterative Boltzmann inversion iteratively constructs potentials $V_i(r)$ with

$$V_{i+1}(r) = V_i(r) + k_B T \ln \frac{g^i(r)}{g(r)} \quad (2)$$

which reproduce $g(r)$ at one single temperature. We will therefore present a simple method of how RDFs can be decomposed as in eq 1, so that in a further step the fitting procedure (eq 2) can be expanded to temperature-dependent RDFs over a large temperature range. Other applications of the SVD in signal processing⁴ and even in biology⁵ have been reported.

The aim of the singular value decomposition is the following. Supposing we have a real matrix $\mathbf{A} = (a_1, a_2, \dots, a_n) \in \mathbb{R}^{m \times n}$ with $m, n \in \mathbb{N}$ and $m \geq n$, i.e., $a_i = (a_{1i}, \dots, a_{mi})^T \in \mathbb{R}^m$ for all $i = 1, \dots, n$. The identification of \mathbf{A} as a compact operator $\mathbf{A}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ between the two Hilbert spaces \mathbb{R}^m and \mathbb{R}^n with the standard scalar product $\langle \cdot | \cdot \rangle$ leads to the application of a well-known theorem from functional analysis.⁶ It states in our case that every matrix \mathbf{A} is a product of three matrices, $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$, and the transpose of $\mathbf{V} \in \mathbb{R}^{n \times n}$:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (3)$$

The matrices $\mathbf{U} = (u_1, u_2, \dots, u_n)$ and $\mathbf{V} = (v_1, v_2, \dots, v_n)$ are orthogonal, and $\mathbf{\Sigma}$ is a diagonal matrix:

$$\mathbf{\Sigma} = \text{diag}(s_1, \dots, s_n) = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & s_n \end{pmatrix} \quad (4)$$

with $s_1 \geq s_2 \geq \dots \geq s_n \geq 0$. This is only one out of several equivalent formulations, but the most useful for our purpose. The s_i are called singular values of \mathbf{A} . In this decomposition, the singular values of the matrix \mathbf{A} are unique. However, the matrices \mathbf{U} and \mathbf{V} are not unique. If there are some degenerate singular values, for instance $s_1 = s_2$, then the vectors u_1, u_2, v_1 , and v_2 are unique up to an orthogonal transformation. In the case of no degeneracy, i.e., $s_1 > s_2 > \dots > s_n \geq 0$, all vectors are unique up to their sign, i.e., an orthogonal transformation on a one-dimensional subspace. However, the nondegeneracy is mostly the case, especially in applications.

From a geometrical point of view, $\mathbf{U}\mathbf{\Sigma}$ as well as \mathbf{A} map the $(n - 1)$ -dimensional unit sphere

$$\begin{aligned} S^{n-1} &:= \{x = (x_1, \dots, x_n)^T \in \mathbb{R}^n \mid \|x\|_2^2 \\ &= x_1^2 + \dots + x_n^2 = 1\} \end{aligned} \quad (5)$$

Received: May 18, 2011

Published: August 31, 2011

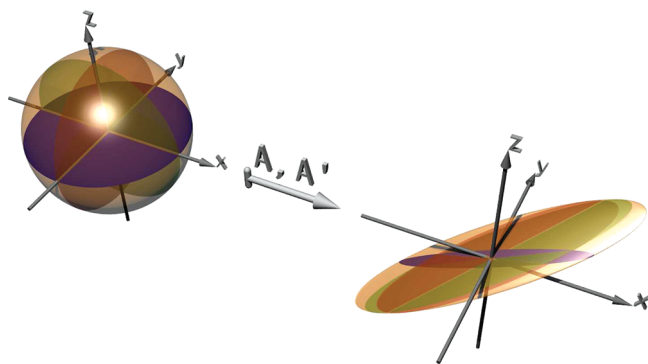


Figure 1. Demonstration and geometrical interpretation of the singular value decomposition (SVD) of two matrices $\mathbf{A} = (a_1, a_2)$ and $\mathbf{A}' = (a_1, a_2, a_3)$ with $a_1 = (0.8, 0.9, 0.6)^T$, $a_2 = (0.4, 0.7, 0.2)^T$, and $a_3 = (0.1, 0.6, 0.0)^T$ as an example of two maps. On the left, the blue disk represents the unit sphere S^1 in the x - y plane, and the orange sphere is the unit sphere S^2 . On the right, both unit spheres are deformed by the maps \mathbf{A} and \mathbf{A}' , respectively, with the singular values $s_1 \approx 1.55770$, $s_2 \approx 0.20571$, $s'_1 \approx 1.64630$, $s'_2 \approx 0.39792$, and $s'_3 \approx 0.03664$.

to an ellipsoid

$$\begin{aligned} E_{\mathbf{A}}^{n-1} &= \{x_1 s_1 u_1 + \dots + x_n s_n u_n \mid x_1^2 + \dots + x_n^2 = 1\} \\ &= \{x_1 a_1 + \dots + x_n a_n \mid x_1^2 + \dots + x_n^2 = 1\} \subset \mathbb{R}^m \end{aligned} \quad (6)$$

This is demonstrated by two matrices \mathbf{A} and \mathbf{A}' in Figure 1 with $m = 3$ and $n = 2$ or 3 . In Figure 1, the unit sphere S^1 (blue disk, left) is mapped to a one-dimensional ellipsoid in \mathbb{R}^3 (blue disk, right), and S^2 is mapped to a two-dimensional ellipsoid (both orange). As seen from this simple illustration, we find that the vectors u_i of \mathbf{U} are the semiaxes of the ellipsoid $E_{\mathbf{A}}^{n-1}$ and the singular values s_i are their length. From the definition of $E_{\mathbf{A}}^{n-1}$, we find the relations

$$s_1 = \max_{x \in E_{\mathbf{A}}^{n-1}} \|x\|_2 \text{ and } s_n = \min_{x \in E_{\mathbf{A}}^{n-1}} \|x\|_2 \quad (7)$$

i.e., the greatest and the smallest distance from the origin.

If we compare both matrices \mathbf{A} and \mathbf{A}' , only the vector a_{n+1} is “added” to \mathbf{A} ($n = 2$), i.e., $\mathbf{A}' = (\mathbf{A}, a_{n+1}) = (a_1, a_2, \dots, a_{n+1})$. The SVD of \mathbf{A}' gives three other matrices \mathbf{U}' , $\mathbf{\Sigma}'$, and \mathbf{V}' with $\mathbf{A}' = \mathbf{U}'\mathbf{\Sigma}'(\mathbf{V}')^T$. In general, all vectors of the primed matrices are different from the nonprimed vectors. For instance, see Figure 1, where the singular values s_1 and s_2 of the blue one-dimensional ellipsoid on the right differ from the three singular values s'_1 , s'_2 , and s'_3 of the orange two-dimensional ellipsoid. However, we have the relation

$$E_{\mathbf{A}}^{n-1} \subseteq E_{\mathbf{A}'}^n \quad (8)$$

which leads to the relation between the largest singular value s_1 of \mathbf{A} and s'_1 of \mathbf{A}' :

$$s'_1 = \max_{x \in E_{\mathbf{A}'}^n} \|x\|_2 \geq \max_{x \in E_{\mathbf{A}}^{n-1}} \|x\|_2 = s_1 \quad (9)$$

as well as the smallest singular value s_n of \mathbf{A} and s'_{n+1} of \mathbf{A}' :

$$s'_{n+1} = \min_{x \in E_{\mathbf{A}'}^n} \|x\|_2 \leq \min_{x \in E_{\mathbf{A}}^{n-1}} \|x\|_2 = s_n \quad (10)$$

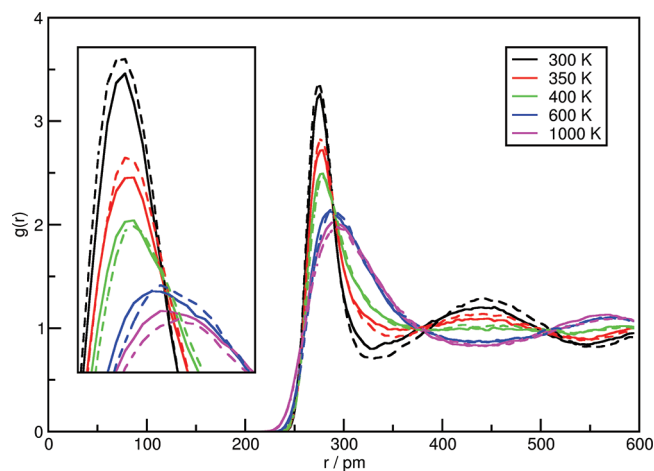


Figure 2. Two sets of temperature-dependent O–O radial distribution functions of water from a previous study⁷ used as examples in the present study. The solid line corresponds to the ruthenium(VI) ester simulation and the dashed line to the ruthenium(VIII) ester simulation.

Table 1. Singular Values s_i from the Decomposition of the Raw RDFs in Figure 3 for both Water Simulations

	simulation sets	
	Ru(VI) ester	Ru(VIII) ester
s_1	29.12231	29.07127
s_2	4.17852	4.87964
s_3	0.41898	0.61054
s_4	0.18050	0.19100
s_5	0.13057	0.10789

Both are useful relations when additional data are used for the analysis.

In order to apply the singular value decomposition to temperature- and pressure-dependent radial distribution functions, the RDFs led to the construction the matrix \mathbf{A} . Each RDF is a single vector, i.e., column $a_i = g(r, T_i)$ of \mathbf{A} . The matrix has the form $\mathbf{A} = (g(r_{ij}, T_j))_{i=1, \dots, m; j=1, \dots, n}$ ($m \geq n$). To demonstrate this, we use the oxygen–oxygen RDFs of water from a previous temperature-dependent ab initio molecular dynamics simulation⁷ where $m = 199$ and $n = 5$, see Figure 2. Two series of five ab initio molecular dynamics simulations (CP2k,⁸ Nosé–Hoover chain thermostat,⁹ BLYP-D¹⁰) were performed with two different ruthenium esters together with 60 water molecules with the temperature ranging from 300 to 1000 K at a constant volume (1203.8 and 1223 pm box length, i.e., $\rho = 1.212$ and 1.170 g cm⁻³; the ruthenium esters increase the density). The simulation time for each temperature was at least 20 ps. The trajectories were analyzed with TRAVIS.¹¹ For the singular value decomposition, we used the algorithm published by Golub and Reinsch in 1970.¹² Additionally, the singular values s_i are sorted in descending order, and the signs of the vectors $u_i = (u_{1i}, \dots, u_{mi})^T$ and v_i were changed to satisfy

$$\sum_{j=1}^m u_{ji} \geq 0 \quad (11)$$

for all $i = 1, \dots, n$. This was done in order to obtain maximal positive u_i .

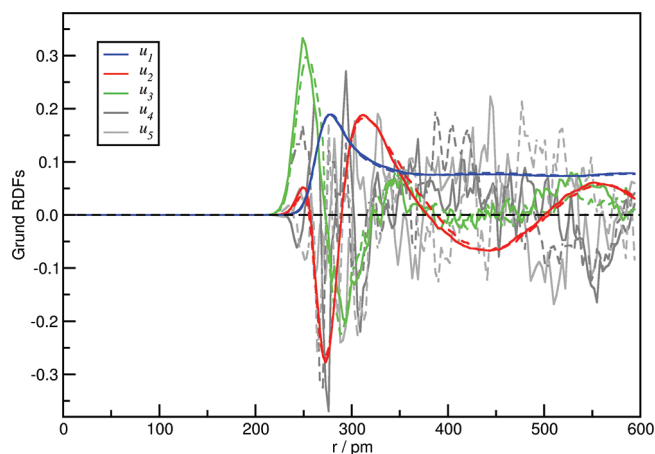


Figure 3. Resulting Grnd RDFs u_1 – u_5 from the singular value decomposition of the raw RDFs from Figure 2 with the corresponding singular values s_i in Table 1.

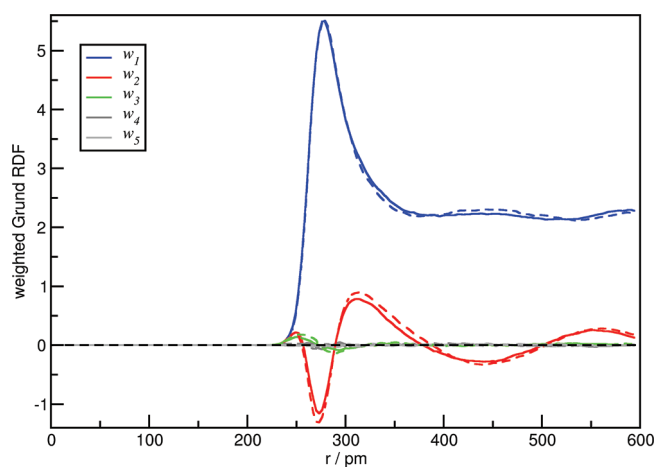


Figure 4. Weighted Grnd RDFs w_1 – w_5 from the singular value decomposition of the raw RDFs from Figure 2.

Applying the SVD to our RDFs, we find the singular values s_i in Table 1. Comparing both simulations, their corresponding singular values are similar. In both cases, the first two singular values are well separated from the last three, which are at least 1 order of magnitude smaller. Because \mathbf{U} and \mathbf{V} are orthogonal, their vectors u_i and v_i are normalized, and the singular values s_i represent the weighting factors in the linear combinations gaining the raw RDFs. Because of the small values of s_3 , s_4 , and s_5 compared to s_1 and s_2 , the raw RDFs in Figure 2 are essentially linear combinations of only two main contributions. The lower singular values only represent noise, which can be filtered out by our approach. Furthermore, from eq 9, we find that with an increasing number of RDFs (spectra), the largest singular value never decreases, while from eq 10, we find that the smallest singular value never increases. Therefore, with an increasing number of RDFs or other data sets, the signal to noise ratio becomes better in contrast to traditional interpretations, where an increased number of spectra only leads to problems in the graphical representation and the noise can only be reduced by better statistics, i.e., longer simulation time for each spectra.

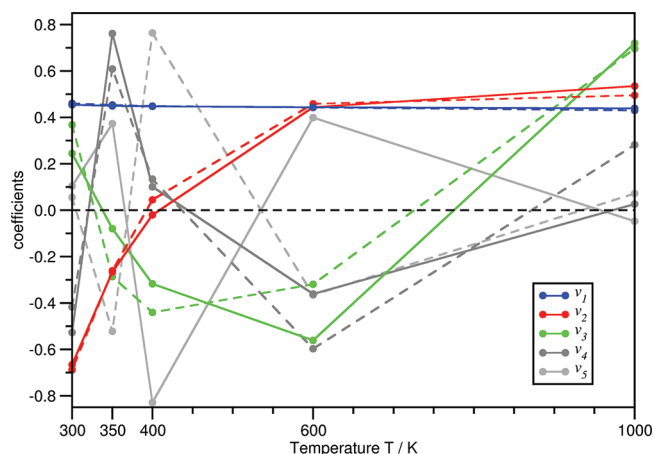


Figure 5. Temperature-dependent development of the coefficients v_1 – v_5 in \mathbf{V} for both simulation sets (Ru(VI) simulation, solid lines; Ru(VIII) simulation, dashed lines).

The Grnd RDFs are the columns u_i of \mathbf{U} , see Figure 3, while the weighted GRDFs (Figure 4) are the columns of $\mathbf{W} = \mathbf{U} \cdot \Sigma$, i.e., $w_i = s_i \cdot u_i$, and they represent the different contributions in a clearer way. In both representations, the first three GRDFs with the highest singular values are colored, while the GRDFs with the smaller singular values (minor GRDFs) are gray. In our application to the RDFs at five temperatures, we observe that the main GRDF (highest singular value of approximately 29) is very smooth and represents the water structure where no second hydration shell occurs. The second GRDF represents the main contributions to the temperature change at a constant density of the raw RDFs. Essentially, all information about the temperature dependence of all five raw RDFs are condensed into one (weighted) GRDF (the second one) and its temperature-dependent coefficients. The third GRDF seems to be a fine structure for the temperature dependence affecting only the first hydration shell up to 325 pm, see Figure 4. At larger distances, it contains much noise. Obviously, this needs further investigation because of the insufficient statistics. In our application, the last two GRDFs are considered statistical noise.

The temperature-dependent coefficients of the GRDFs (columns v_i of \mathbf{V}) are shown in Figure 5. The coefficients v_1 of the main GRDF (blue) change slightly over the whole temperature range from 300 to 1000 K. Therefore, in our application, the main GRDF seems to be a constant contribution to all water structures. A systematic change of the coefficients is found for the second GRDF, see v_2 (red curve) in Figure 5. With increasing temperature, the coefficients rise from ~ -0.7 at 300 K up to ~ 0.6 at 1000 K. This behavior is in accordance with the previous interpretation that the main temperature changes are condensed in the second GRDF. The second GRDF has negative values in the region of the second hydration shell. Values of v_2 smaller than zero characterize a phase with a second hydration shell for the water molecules. But with increasing temperature, v_2 increases and characterizes a vanishing second hydration shell. In the general case, the coefficients and GRDFs are temperature- as well as density-dependent, as seen from the comparison in Figure 5. For a better interpolation, additional temperatures are needed. In general, interpolation of the coefficients at temperatures

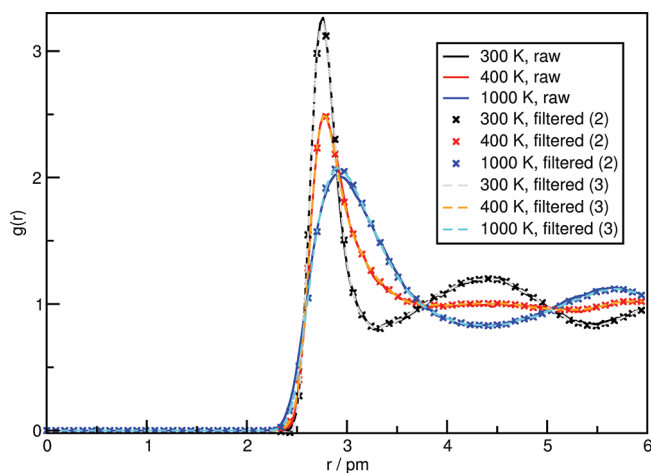


Figure 6. Reconstruction of three raw RDFs (Figure 2) at 300, 400, and 1000 K using only the contributions from the two or three first GRDFs and without the minor contributions from the remaining GRDFs (noise).

between the simulated temperatures is feasible and leads to the possibility of predicting all RDFs in a given temperature range and therefore predicting all averages depending on $g(r)$:¹³

$$\begin{aligned} \langle A \rangle &= 2\pi N\rho \int_0^\infty a(r) g(r) r^2 dr \\ &= \sum_i v_i(T) \cdot 2\pi N\rho \int_0^\infty a(r) w_i(r) r^2 dr = \sum_i v_i(T) \cdot A_i \end{aligned} \quad (12)$$

as well as their temperature derivatives. Additionally, the irregular temperature development of v_3 , v_4 , and especially v_5 is in accordance with its interpretation as noise from the statistics. The reconstruction of the temperature-dependent RDFs without the noise from u_3 , u_4 , and u_5 is shown in Figure 6. The reconstructed RDFs are not changed significantly in comparison to the raw RDFs in Figure 2, where the noise (u_4 , u_5) is left out. At 300 and 400 K, the raw and filtered RDF coincide over the full range, while at 1000 K the missing contribution from the fine structure (u_3) slightly affects the first hydration shell. The contribution from the fine structure u_3 restores the first hydration shell at 1000 K, and the raw and reconstructed RDFs coincide over the whole range. As seen from the RDFs in Figure 2, our approach is not restricted by describing only the temperature changes accurately, because similar changes of the RDFs (or even other functions like autocorrelation functions etc.) are also found for many other continuous dependencies, and SVD can therefore be applied.

In summary, we presented a simple, systematic, and unified approach for analyzing temperature- and pressure-dependent radial distribution functions and structures from the singular value decomposition. Its generality is also suited for the analysis of other dependencies and functions. Furthermore, it could be useful for the comparison of molecular dynamics methods like force field techniques with different force fields and ab initio techniques with different DFT functionals, especially by explaining temperature- and pressure-induced structural changes. The main advantages of our approach are that an increasing number of RDFs systematically improves the results, as seen from eqs 9 and 10, and that the comparison and the analysis of a

large number of RDFs (spectra) are easier because all RDFs are reduced to some main GRDFs and the minor GRDFs can be left out.

The main GRDFs contain the main contributions and are easier to handle. The minor GRDFs are found to contain the statistical noise only, which can be filtered out by our approach. We investigated five raw oxygen–oxygen RDFs of water from ab initio molecular dynamics simulations and found that all RDFs are essentially a linear combination of only two GRDFs. The temperature-dependent problem reduced therefore from five RDFs (spectra) to only two GRDFs (spectra) and their temperature-dependent coefficients. The remaining three degrees of freedom are found to be statistical noise, which was filtered out without significantly changing the resulting raw RDFs. In AIMD simulations, the simulation time is very limited. Unfortunately, this results in unsatisfactory statistics, a major drawback for extensive temperature studies with AIMD simulations. This drawback is overcome by our approach because the number of spectra and, therefore, the overall simulation time (not only the simulation time at each temperature) improves the results and the statistics, and intermediate temperatures can be interpolated. Therefore, our approach reveals new ways of investigating and interpreting the temperature and pressure dependence of molecular systems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: didio@uni-leipzig.de.

ACKNOWLEDGMENT

This work was supported by the DFG, in particular by the project KI 768/6-1, graduate school BuildMoNa, and the ESF. Computer time from the RZ Leipzig was gratefully acknowledged.

REFERENCES

- (1) (a) Frauenfelder, H.; Petsko, G. A.; Tsernoglou, D. *Nature* **1979**, *280*, 558–563. (b) Bertagnolli, H. *Angew. Chem., Int. Ed.* **1992**, *31*, 1577–1578. (c) Corongiu, G.; Clementi, E. *J. Chem. Phys.* **1992**, *97*, 2030–2038. (d) Laasonen, K.; Sprik, M.; Parrinello, M.; Car, R. *J. Chem. Phys.* **1993**, *99*, 9080–9089. (e) Mancera, R. L.; Buckingham, A. D. *Chem. Phys. Lett.* **1995**, *234*, 296–303. (f) Mancera, R. L.; Buckingham, A. D. *J. Phys. Chem.* **1995**, *99*, 14632–14640. (g) Sprik, M.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1996**, *105*, 1142–1152. (h) Yoshii, N.; Yoshie, H.; Miura, S.; Okazaki, S. *J. Chem. Phys.* **1998**, *109*, 4873–4884. (i) Robinson, G. W.; Cho, C. H.; Urquidi, J. *J. Chem. Phys.* **1999**, *111*, 698–702. (j) Sorenson, J. M.; Hura, G.; Glaeser, R. M.; Head-Gordon, T. *J. Chem. Phys.* **2000**, *113*, 9149–9161. (k) Handgraaf, J.-W.; van Erp, T. S.; Meijer, E. *J. Chem. Phys. Lett.* **2003**, *367*, 617–624. (l) Hura, G.; Russo, D.; Glaeser, R. M.; Head-Gordon, T.; Krack, M.; Parrinello, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1981–1991. (m) van Erp, T. S.; Meijer, E. *J. Chem. Phys.* **2003**, *118*, 8831–8840. (n) Mantz, Y. A.; Chen, B.; Martyna, G. *J. Chem. Phys. Lett.* **2005**, *405*, 294–299. (o) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2006**, *125*, 154507:1–14. (p) Lehmann, S. B. C.; Spickermann, C.; Kirchner, B. *J. Chem. Theory Comput.* **2009**, *5*, 1640–1649. (q) Lehmann, S. B. C.; Spickermann, C.; Kirchner, B. *J. Chem. Theory Comput.* **2009**, *5*, 1650–1656. (r) Guse, C.; Simionescu, A.; Schünemann, B.; Hentschke, R. *J. Phys.: Condens. Matter* **2010**, *22*, 325105:1–14. (2) (a) Urquidi, J.; Singh, S.; Cho, C. H.; Robinson, G. W. *Phys. Rev. Lett.* **1999**, *83*, 2348–2350. (b) Head-Gordon, T.; Hura, G. *Chem. Rev.* **2002**, *102*, 2651–2670.

- (3) Carbone, P.; Varzaneh, H. A. K.; Chen, X.; Müller-Plathe, F. *J. Chem. Phys.* **2008**, *128*, 064904:1–11.
- (4) (a) Biglieri, E.; Yao, K. *Signal Process* **1989**, *18*, 277–289. (b) Edfors, O.; Sandell, M.; van de Beek, J.-J.; Wilson, S.; Borjesson, P. *IEEE Trans. Commun.* **1998**, *46*, 931–939.
- (5) (a) Alter, O.; Brown, P. O.; Botstein, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10101–10106. (b) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. *Bioinformatics* **2001**, *17*, 520–525.
- (6) (a) Reed, M.; Simon, B. *Methods of Modern Mathematical Physics: Functional Analysis*; Academic Press: San Diego, CA, 1980; Vol. 1, Theorem VI.17. (b) Werner, D. *Funktionalanalysis*, 6th ed.; Springer-Verlag: Berlin, Germany, 2007; Satz VI.3.6.
- (7) di Dio, P. J.; Brehm, M.; Kirchner, B. Temperature Dependent Hydration of Ruthenium(VI)-dioxo-2,5-dioxolane and Ruthenium(VIII)-trioxo-2,5-dioxolane: An Ab Initio Molecular Dynamics Study. Submitted.
- (8) (a) Lippert, G.; Hutter, J.; Parrinello, M. *Theor. Chem. Acc.* **1999**, *103*, 124–140. (b) CP2k developers group under the terms of the GNU General Public License. See: CP2P Developers Home Page. <http://cp2k.berlios.de/> (accessed Jul 26, 2011).
- (9) (a) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519. (b) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697. (c) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *J. Chem. Phys.* **1992**, *97*, 2635–2643.
- (10) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789. (c) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473. (d) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (11) Brehm, M.; Kirchner, B. *J. Chem. Inf. Model.* **2011**, *51*, 2007–2023.
- (12) Golub, G. H.; Reinsch, C. *Numer. Math.* **1970**, *14*, 403–420.
- (13) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, Great Britain, 1987.

Locality and Fluctuations: Trends in Imidazolium-Based Ionic Liquids and Beyond

Katharina Wendler,[†] Stefan Zahn,[‡] Florian Dommert,[§] Robert Berger,^{||} Christian Holm,[§] Barbara Kirchner,[‡] and Luigi Delle Site^{*,†}

[†]Max-Planck-Institut für Polymerforschung, Ackermannweg 10, D-55128 Mainz, Germany

[‡]Universität Leipzig, Linnéstraße 2, D-04103 Leipzig, Germany

[§]Institut für Computerphysik, Universität Stuttgart, Pfaffenwaldring 27, D-70569 Stuttgart, Germany

^{||}Clemens-Schöpf-Institut, Technische Universität Darmstadt, Petersenstrasse 22, D-64287 Darmstadt, Germany

ABSTRACT: Three different imidazolium-based ionic liquids, 1,3-dimethylimidazolium chloride, 1-ethyl-3-methylimidazolium thiocyanate, and 1-ethyl-3-methylimidazolium dicyanamide, are investigated by Car–Parrinello simulations. A common behavior, such as a broad electric dipole moment distribution of the ions and a related high degree of locality, is found to characterize all these systems. Going beyond imidazolium-based systems, we found that even for the protic ionic liquid monomethyl ammonium nitrate, the same features hold. These results represent a strong support to the hypothesis of rattling ions in long-living ion cages proposed in the last years.

1. INTRODUCTION

Ionic liquids (ILs) are defined as low-melting salts being often liquid at room temperature. Due to their unique physicochemical properties, ILs have gained immense interest.^{1,2} For instance, they offer tailor-made high-potential solutions in chemical reactions as catalysts,³ in separation processes due to their adaptable solubility,⁴ or being ionic, as electrolytes.⁵ Moreover, ILs dissolve cellulose and permit to reshape and process the most abundant organic raw material using less energy and chemicals and increasing the processes sustainability.⁶ Chemically, ILs are designed and synthesized by combining two ions; however, a reasonable accurate prediction of newly formed ILs' properties, employing basic theoretical principles, is often not possible, as the underlying physics is not yet fully understood.⁷ Current theoretical approaches and (above all) large-scale simulation methods, being based mostly on empirical models, can provide only a rather limited capability of predictions of properties and a partial interpretation of experimental data.^{8,9} Thus, there is a vivid interest in elucidating the governing interactions, Coulomb and van der Waals forces¹⁰ as well as hydrogen bonding^{11–18} and their subtle interplay. As for any molecular liquid, this interplay is expressed by the connection between the local scale, i.e., how a molecule reacts to its immediate surrounding in the short time, and the global mesoscopic scale, i.e., the bonding network or diffusion properties which typically occur on a larger scale.^{9,19} In this context, for this work we intend to identify basic universal physical principles for (at least a class of) ILs. To this aim, we study a series of imidazolium-based ILs, 1,3-dimethylimidazolium chloride (MMIM Cl), 1-ethyl-3-methylimidazolium thiocyanate (EMIM SCN), and 1-ethyl-3-methylimidazolium dicyanamide (EMIM DCA) as well as the protic IL monomethyl ammonium nitrate (MMAN) (see Figure 1) using state of the art electronic structure methods and proceed to a systematic comparison among them. In previous work on MMIM Cl,¹⁹ the electric dipole moment distribution was found to be very much

spread, indicating a strong degree of fluctuation, but the electrostatic properties were rather local. Further studies confirmed the strong electrostatic screening (theoretical,^{9,10,20,21} experimental²²). The natural question arising is whether this behavior is due to any peculiar chemical structure of the ions or may be a sign of a more general behavior. Here, we show that these properties are common to several ionic liquids, and thus, they may represent a “true ionic liquid” signature.²²

2. COMPUTATIONAL DETAILS

This study makes use of some of the largest ab initio simulations of ILs ensuring sampling and equilibration so far not reached. We used the Car–Parrinello approach implemented in the CPMD code^{23,24} with the BLYP functional.^{25,26} We employ Troullier–Martins pseudopotentials, derived using the BLYP functional, and an empirical dispersion correction,²⁷ which has been accurately tested in previous work.²⁸ Further checks were done for this work by studying three one-ion pair clusters and comparing structures and geometries with those obtained by high-level quantum chemical methods. For the MMAN ion pairs, we used instead the DCACP-BP pseudopotentials,²⁹ which were tested accurately for this system in previous work.¹⁴ The simulations were started from a snapshot of a classical MD simulation, and the equilibration time was chosen to be 5 ps unless otherwise specified. A short overview of the systems is given below: (1) 8 ion pairs EMIM SCN, 400 K, (12.64 Å)³, 67.66 ps; (2) 32 ion pairs EMIM SCN, 400 K, (20.06 Å)³, 6.24 ps (started from an ab initio trajectory¹¹ snapshot); (3) 48 ion pairs EMIM SCN, 400 K, (22.97 Å)³, 1.0 ps (0.5 ps equilibration); (4) 8 ion pairs EMIM DCA, 400 K, (13.11 Å)³, 63.50 ps; (5) 30 ion pairs EMIM DCA, 400 K, (20.37 Å)³, 43.85 ps; (6) 64 ion pairs EMIM DCA, 400 K, (26.23 Å)³, 4.5 ps. (0.5 ps equilibration); (7) 8 ion pairs MMAN,

Received: June 6, 2011

Published: September 01, 2011

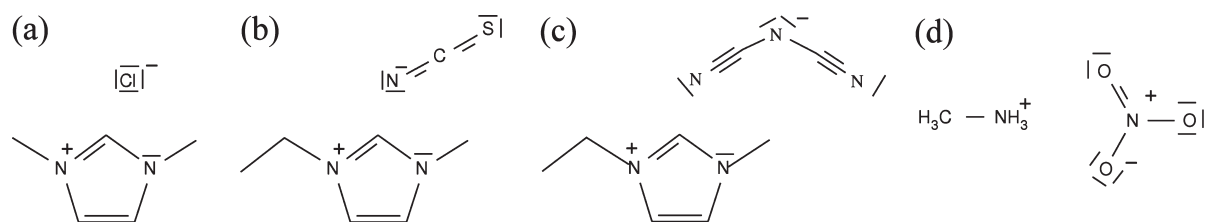


Figure 1. Presentation of one mesomeric structure of (a) MMIM Cl, (b) EMIM SCN, (c) EMIM DCA, and (d) MMAN.

400 K, $(10.19 \text{ \AA})^3$, 84.02 ps; (8) 48 MMAN ion pairs where 100 snapshots were randomly selected for analysis from the trajectory of ref 14; (9) 8 ion pairs MMIM Cl, 425 K, $(11.62 \text{ \AA})^3$, 79.66 ps. The latter was done with the same setup as the corresponding larger system of 30 MMIM Cl pairs.¹⁹ The two larger systems of 48 ion pairs of EMIM SCN and 64 ion pairs of EMIM DCA were used here as preliminary tests to indicate possible size effects on the distributions of the dipoles and on the correlations studied. The large computational effort required allows only for short simulations, and technically, it is important to justify why one may use a short equilibration time in this context. The starting configurations obtained from the equilibration by classical simulations gave interatomic forces which were smaller than 0.003 au after the first step of the quantum calculation. This implied that the starting configuration provided a reasonable overall liquid structure and that the equilibration time chosen (0.5 ps) was sufficient to equilibrate the local electronic degrees of freedom. As a further test, we considered 10 uncorrelated classical configurations to sample statistically the liquid structure found in the classical molecular dynamics (MD). After performing a wave function optimization, the largest force acting on an atom was below 0.004 au for each of them. In a similar study of ionic liquids in literature, a much larger value of 0.1 au was taken for identifying equilibrated (at quantum level) configurations obtained from classical samples.³⁰ The resulting electron densities were used to calculate dipoles and their distributions; we found the same results obtained for smaller systems. Since we did not consider quantities linked directly to the dynamic properties of the system (for which a long equilibration is mandatory), the technical set up used for these larger calculations was sufficient to represent a preliminary test. The dipole moments were calculated using the maximally localized Wannier analysis^{31,32} every 500th time step. In general, density functional theory (DFT) and Wannier center-based dipole moments were found to be consistent with experimental values for single molecules in the gas phase.^{33,34} For 100 snapshots per system, the Blöchl analysis was done as in ref 9.

3. RESULTS AND DISCUSSION

3.1. Molecular Dipole Moment As Physical Indicator.

In general, the electric dipole moment of an ion (denoted as dipole in the following), contrary to that of a neutral molecule, is not defined uniquely, as it depends on the arbitrary choice of the origin with respect to which it is calculated. This implies that the absolute value of ionic dipoles does not have a physical quantitative meaning but, defined in a consistent way for all systems, can be employed for a systematic comparison among different ions to identify some generic trends. Here, the dipole moment of imidazolium ions is calculated with respect to the geometric center of the five ring atoms (COR), the dipole moment of all

other ions with respect to the geometric center of all atoms (COG). However, our findings are largely independent of such a choice. Molecular dipoles in ILs are also not experimentally measurable, however, they represent a powerful tool for a theoretical interpretation of the system properties. In fact, the ion dipole moment is an indicator for charge displacement as well as (being a vector) local molecular packing. It sums up electronic and steric influences and describes the interplay of charge and shape; hence, it represents an accurate descriptor for the variation of the ion's electrostatic properties upon local liquid configurational changes. Moreover, the evaluation of dipole correlations along the whole system provides information about the effective range of electronic interactions.

3.2. Fluctuations. Figure 2 reports the dipole distributions of anions and cations for all the systems considered. In general, the distributions are broader compared to those of nonionic liquids, such as liquid water.³³ The MMIM and EMIM cation dipole distributions have about the same mean value with a strongly pronounced spread in all three cases essentially independent of the counterion. For the chemically more different anions of the imidazolium ILs, a trend emerges: given the size of each specific anion, the spread of the distributions is also very pronounced and rises as the size and polarizability of anions increase. In fact, a larger size can imply a larger number of both conformational degrees of freedom (e.g., internal vibrations and local motion of the side chain) and electronic degrees of freedom (i.e., possibility of polarization). We may conclude that all these imidazolium systems are characterized by large electrostatic fluctuations and that the difference between the imidazolium systems is dictated mostly by the nature of the anion and not so much by the chemical specificity of MMIM or EMIM. Going beyond imidazolium-based systems to the protic IL MMAN, we have found that this general trend still holds although in reverse order, i.e., the anion has a broader dipole distribution than the cation, which can be motivated by the larger van der Waals volume of the anion. Interestingly, the gas-phase polarizability of the nitrate anion is significantly lower than the ones of the SCN or DCA anions, but the nitrate dipole distribution has a larger spread. This suggests that, in general, all these systems are characterized by strong electrostatic fluctuations. The natural question arising at this point is what the origin of these fluctuations is: Are they the product of bulk density fluctuations on a large length scale, or are they the result of very local but largely diverse molecular packing?

A comparison with the results obtained with classical flexible models shows that the spread of the dipole distribution of the quantum chemical calculations is 50% larger than that of the classical studies. Furthermore in quantum chemical calculations, the ions dipole moment distributions of a gas-phase ion pair are also broad. The spread is larger compared to the liquid phase. This is mainly due to the large number of ion pair configurations associated with strong mutual polarization. The results above

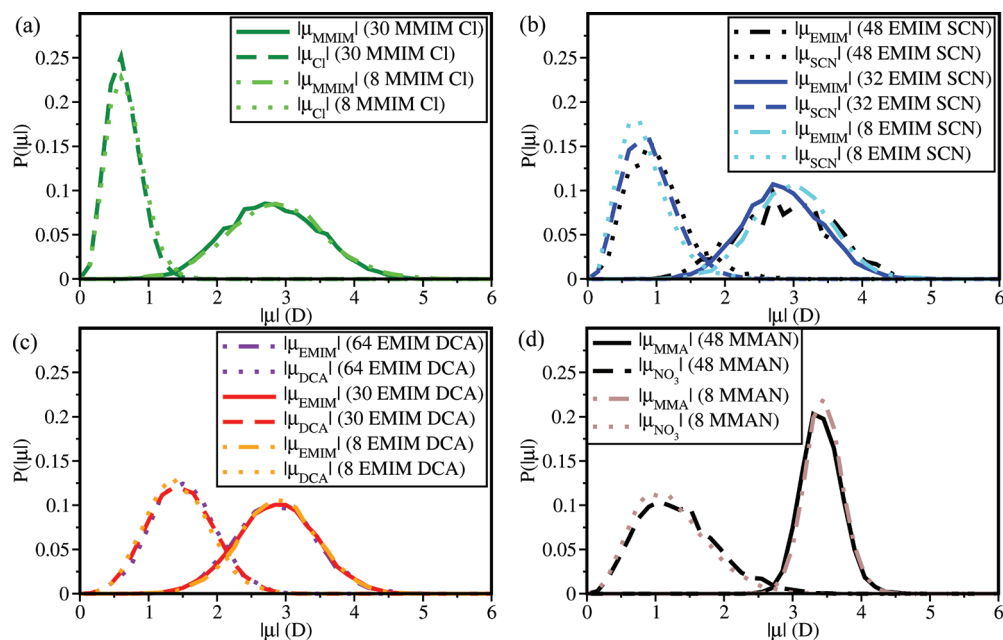


Figure 2. Dipole moment distributions for large and small (≥ 30 and 8 ion pairs, respectively) IL systems of (a) MMIM Cl, (b) EMIM SCN, (c) EMIM DCA, and (d) MMAN.

indicate that while, in general, the large spread seems intrinsic to ILs and their electronic properties, the actual dipole moment distributions are very sensitive to the specific ions environment. The decisive interactions are characterized by a certain range that needs to be quantified.

In the next section, we address this point by comparing various correlation functions and by comparing large systems with smaller systems of eight ion pairs.

3.3. Locality. Figure 2 reports the dipole distributions for both the large systems considered and their thermodynamical equivalent systems of eight ion pairs. It is remarkable that the ion dipole moment distributions of the large systems and the eight ion pair systems are the same in all cases. A possible scenario emerging from this comparison is that an ion is influenced only by its immediate neighbors in the short time and seems to not experience perturbations from the long-range behavior of the liquid. Indeed, the counterions in the first shell overcompensate the charge of an ion, thus, showing overscreening.^{20,35,36} This result would strongly support the idea of ions rattling in a long-living ion cage proposed in literature.^{14,37–40} The statement above is certainly true for the simulations done here. However, at this stage, it should not be taken as a general claim but only as an indication of a possible trend. In fact, the box sizes of the systems providing a statistically significant set of reliable data (i.e., ion pairs < 48) are truncated at distances well short of those where the radial distribution functions go to unity. It is, however, rather encouraging that short simulation tests on larger systems (48 ion pairs EMIM SCN and 64 ion pairs EMIM DCA), which do not suffer from the limitation above, confirm all the conclusions reached here. To further check the idea of locality, we test the range of extension of three types of electrostatic interactions: monopole–monopole, monopole–dipole, and dipole–dipole. The interionic radial distribution functions would not capture the electrostatic effects, as they are dominated by excluded volume interactions. To obtain the monopole interactions between ions, we fit a set of partial charges on the atom centers that give a best

Table 1. Cation Charges in Units of e Obtained via Blöchl Analysis⁹

system	MMIM Cl	EMIM SCN	EMIM DCA	MMAN
≥ 30	0.63 ± 0.15	0.56 ± 0.25	0.67 ± 0.21	0.55 ± 0.29
8	0.64 ± 0.16	0.56 ± 0.25	0.70 ± 0.19	0.56 ± 0.27

fit to the multipole moments of the bulk ion configurations in Fourier space via the Blöchl method.^{9,41} Surprisingly enough, we find for all investigated systems, ionic charges considerably less than one, in the range 0.55–0.7 in units of the elementary charge e , see Table 1. Force fields with ad hoc reduced ionic charges,^{21,37,42} treating this just as another free parameter in the force field description, have been partially successful to reproduce dynamical quantities of ILs besides the static ones. In this context, our results suggest that the charge reduction may be a real physical effect as proposed by experiments.⁴³ For instance for EMIM DCA, the reduced charges correspond to an electronic dielectric constant $\epsilon_{\text{el}} = (q/q_{\text{eff}})^2 \approx 2.2$, which is in agreement with the experimental refraction index $n = (\epsilon_{\text{el}})^{1/2} \approx 1.5$.⁴⁴ Moreover, the cation charges in the large and the small systems obtained via this analysis are the same within the standard deviation in all cases, demonstrating that this charge reduction effect takes place over a range of only a few angstroms. Thus, eight ion pairs are sufficient to reproduce the effective monopole structure of each ion. For the monopole–dipole correlation, we consider the angular distribution between the dipole of an ion and the vector from the ion's center of reference to another ion's center of charge (COC), e.g., taking one cation per time as reference: $\cos \Phi = \mu_{\text{cat}}^{\text{ref}} \cdot \mathbf{r}_{\text{cat}}^{\text{shell}} / |\mu_{\text{cat}}^{\text{ref}}| \cdot |\mathbf{r}_{\text{cat}}^{\text{shell}}|$; $\mu_{\text{cat}}^{\text{ref}}$ is the dipole of the reference cation, while $\mathbf{r}_{\text{cat}}^{\text{shell}}$ is the vector from the COR of the reference ion to the COC of cations in the i th radial bin ($i = 1$: below 5 Å, $i = 2$: 5–7 Å, $i = 3$: 7–9 Å, $i = 4$: above 9 Å). All possible distributions were calculated: $\mu_{\text{an}} - \mathbf{r}_{\text{an}}$, $\mu_{\text{an}} - \mathbf{r}_{\text{cat}}$, $\mu_{\text{cat}} - \mathbf{r}_{\text{an}}$, and $\mu_{\text{cat}} - \mathbf{r}_{\text{cat}}$. Similarly, all possible dipole–dipole angular distributions as $\mu_{\text{an}} - \mu_{\text{an}}$, $\mu_{\text{cat}} - \mu_{\text{an}}$, and $\mu_{\text{cat}} - \mu_{\text{cat}}$ were

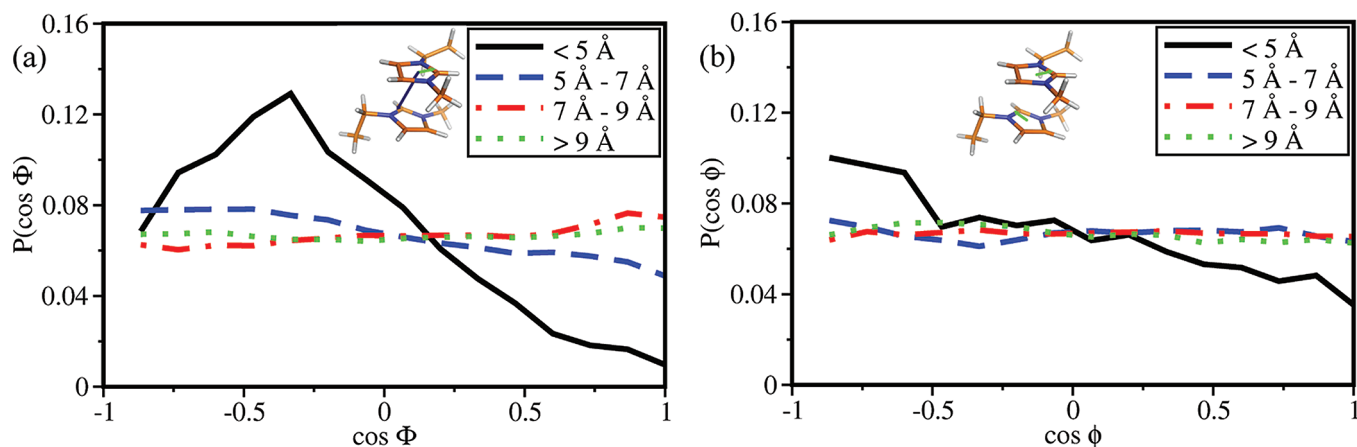


Figure 3. Angular probability distribution of (a) $\cos \Phi$ (charge–dipole) and (b) $\cos \phi$ (dipole–dipole) of cations separated by distances below 5 Å (black), from 5–7 Å (blue), 7–9 Å (red), or above 9 Å (green) in the 30 EMIM DCA system.

studied as in ref 19, e.g.: $\cos \phi = \mu_{\text{cat}}^{\text{ref}} \cdot \mu_{\text{cat}}^{\text{shell}} / |\mu_{\text{cat}}^{\text{ref}}| \cdot |\mu_{\text{cat}}^{\text{shell}}|$ in which $\mu_{\text{cat}}^{\text{shell}}$ is the dipole of cations in the i th radial bin. As examples, $\mu_{\text{cat}} - \mathbf{r}_{\text{cat}}$ and $\mu_{\text{cat}} - \mu_{\text{cat}}$ of the 30 EMIM DCA system are given in Figure 3. Basically, the dipole of a cation correlates only in the very immediate neighborhood with both the direction to the COC and the dipole of cations, while displaying no preferential alignment beyond this short range. All other distributions give the same qualitative picture. For the cation–anion and the anion–anion dipole–dipole angular distributions, only a negligible preferential order is observed even in the first radial bin, while the correlations of the dipoles and the COC directions decay slower. In any case at distances above 8 Å, there is no preferential alignment in any of the distributions. Given the diversity in the chemical structure of the ions of each system and the fact that the short simulation tests on larger systems confirm the results, we are tempted to propose that the locality and the fluctuations may be a general characteristic of, at least, a large class of ILs. The question arising at this point is what happens at the local level that produces these pronounced fluctuations? We can provide some hints for the imidazolium-based case.

We have found that the dipole moment distribution of ions depends neither on the overall numbers of close cations or anions nor on the magnitude of dipole moments of adjacent ions. However, there are some geometries¹⁵ that may, at least in part, clarify the molecular origin of the dipole fluctuations. We have identified two counterbalancing aspects: ring stacking that lowers dipole moments and hydrogen bonding that increases them.^{45,46} Although the complexity is much higher and probably involves several other balancing effects, a basic qualitative explanation of locality and fluctuations is given by the fact that as ions move between environments that are hydrogen bonding dominated to those which are ring stacking dominated (and vice versa), the dipoles change from larger to smaller values (and vice versa) leading to a broad distribution.

4. CONCLUSIONS

We have studied a series of imidazolium-based ILs and, for comparison, one protic IL of different chemical nature. For all systems, we find that static electrostatic properties are the result of two main aspects: locality and fluctuations. Basic considerations about the physics of these systems lead us to propose that these two aspects may actually be a more general signature for a larger class of ILs. Despite the fact that properties of electrolytes

are governed mainly by long-range electrostatic interactions, we found that the electrostatic interactions in the ILs investigated by us were reduced and very local, yet dominated by large fluctuations. In general, one expects screening in a liquid consisting of ions. However, it is not obvious that this takes place on a range comparable to the molecular size. Our results can be useful for a rational design of further molecules and for providing building criteria for large-scale theoretical and simulation methods. Likely, the high locality and charge reduction are the reasons of the success of classical force fields with empirically reduced ion charges^{21,37,42} in describing bulk properties.

AUTHOR INFORMATION

Corresponding Author

*E-mail: dellsite@mpip-mainz.mpg.de.

ACKNOWLEDGMENT

All authors would like to thank the DFG priority program 1191 “Ionic Liquids” for funding, Björn Baumeier for fruitful discussions, Denis Andrienko for critical reading of the manuscript, and the RZG of the Max Planck Society and the HLRS Stuttgart for providing the necessary computer time.

REFERENCES

- (1) Pádua, A. A. H.; Costa Gomes, M. F.; Canongia Lopes, J. N. A. *Acc. Chem. Res.* **2007**, *40*, 1087–1096.
- (2) Plechkova, N. V.; Seddon, K. R. *Chem. Soc. Rev.* **2008**, *37*, 123–150.
- (3) *Ionic Liquids in Synthesis*; Wasserscheid, P., Welton, T., Eds.; Wiley-VCH: Hoboken, NJ, 2007.
- (4) Werner, S.; Haumann, M.; Wasserscheid, P. *Annu. Rev. Chem. Biomol. Eng.* **2010**, *1*, 203–230.
- (5) Armand, M.; Endres, F.; MacFarlane, D. R.; Ohno, H.; Scrosati, B. *Nat. Mater.* **2009**, *8*, 621–629.
- (6) Pinkert, A.; Marsh, K. N.; Pang, S.; Staiger, M. P. *Chem. Rev.* **2009**, *109*, 6712–6728.
- (7) Kirchner, B. *Top. Curr. Chem.* **2009**, *290*, 213–262.
- (8) Dommert, F.; Schmidt, J.; Krekeler, C.; Zhao, Y. Y.; Berger, R.; Delle Site, L.; Holm, C. *J. Mol. Liq.* **2010**, *152*, 2–8.
- (9) Schmidt, J.; Krekeler, C.; Dommert, F.; Zhao, Y.; Berger, R.; Delle Site, L.; Holm, C. *J. Phys. Chem. B* **2010**, *114*, 6150–6155.

- (10) Zahn, S.; Uhlig, F.; Thar, J.; Spickermann, C.; Kirchner, B. *Angew. Chem., Int. Ed.* **2008**, *47*, 3639–3641.
- (11) Thar, J.; Brehm, M.; Seitsonen, A. P.; Kirchner, B. *J. Phys. Chem. B* **2009**, *113*, 15129–15132.
- (12) Tsuzuki, S.; Tokuda, H.; Mikami, M. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4780–4780.
- (13) Lehmann, S. B. C.; Roatsch, M.; Schöppke, M.; Kirchner, B. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7473–7486.
- (14) Zahn, S.; Thar, J.; Kirchner, B. *J. Chem. Phys.* **2010**, *132*, 124506–124506.
- (15) Qiao, B.; Krekeler, C.; Berger, R.; Delle Site, L.; Holm, C. *J. Phys. Chem. B* **2008**, *112*, 1743–1751.
- (16) Dommert, F.; Schmidt, J.; Qiao, B.; Zhao, Y.; Krekeler, C.; Delle Site, L.; Berger, R.; Holm, C. *J. Chem. Phys.* **2008**, *129*, 224501.
- (17) Stark, A. *Top. Curr. Chem.* **2009**, *290*, 41–81.
- (18) Fumino, K.; Wulf, A.; Ludwig, R. *Angew. Chem., Int. Ed.* **2008**, *47*, 3830–3834.
- (19) Krekeler, C.; Dommert, F.; Schmidt, J.; Zhao, Y. Y.; Holm, C.; Berger, R.; Delle Site, L. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1817–1821.
- (20) Lynden-Bell, R. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1733–1740.
- (21) Youngs, T. G. A.; Hardacre, C. *ChemPhysChem* **2008**, *9*, 1548–1558.
- (22) Hallett, J. P.; Liotta, C. L.; Ranieri, G.; Welton, T. *J. Org. Chem.* **2009**, *74*, 1864–1868.
- (23) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- (24) CPMD; IBM and Max-Planck Institut, Stuttgart: Stuttgart, Germany; <http://www.cpmid.org/>.
- (25) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3098.
- (26) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (27) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (28) Zahn, S.; Kirchner, B. *J. Phys. Chem. A* **2008**, *112*, 8430–8435.
- (29) Lin, I.-C.; Röthlisberger, U. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2730–2730.
- (30) Bagno, A.; D’Amico, F.; Saielli, G. *ChemPhysChem* **2007**, *8*, 873–881.
- (31) Wannier, G. H. *Phys. Rev.* **1937**, *52*, 191–197.
- (32) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, 12847–12847.
- (33) Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308–3308.
- (34) Kirchner, B.; Hutter, J. *J. Chem. Phys.* **2004**, *121*, 5133–5142.
- (35) Kornyshev, A. A. *J. Phys. Chem. B* **2007**, *111*, 5545–5557.
- (36) Fedorov, M. V.; Kornyshev, A. A. *J. Phys. Chem. B* **2008**, *112*, 11868–11872.
- (37) Morrow, T. I.; Maginn, E. J. *J. Phys. Chem. B* **2002**, *106*, 12807–12813.
- (38) Del Popolo, M. G.; Voth, G. A. *J. Phys. Chem. B* **2004**, *108*, 1744–1752.
- (39) Turton, D. A.; Hunger, J.; Stoppa, A.; Hefter, G.; Thoman, A.; Walther, M.; Buchner, R.; Wynne, K. *J. Am. Chem. Soc.* **2009**, *131*, 11140–11146.
- (40) Méndez-Morales, T.; Carrete, J.; Cabeza, O.; Gallego, L. J.; Varela, L. M. *J. Phys. Chem. B* **2011**, *115*, 6995–7008.
- (41) Blöchl, P. E. *J. Chem. Phys.* **1995**, *103*, 7422–7422.
- (42) Bhargava, B. L.; Balasubramanian, S. *J. Phys. Chem. B* **2007**, *111*, 4477–4487.
- (43) Cremer, T.; Kolbeck, C.; Lovelock, K. R. J.; Paape, N.; Wölfel, R.; Schulz, P. S.; Wasserscheid, P.; Weber, H.; Thar, J.; Kirchner, B.; Maier, F.; Steinrück, H.-P. *Chem.—Eur. J.* **2010**, *16*, 9018–9033.
- (44) Fröba, A. P.; Kremer, H.; Leipertz, A. *J. Phys. Chem. B* **2008**, *112*, 12420–12430.
- (45) Zahn, S.; Wendler, K.; Delle Site, L.; Kirchner, B. *Phys. Chem. Chem. Phys.* **2011**, *13*, 15083–15093.
- (46) Wendler, K.; Dommert, F.; Zhao, Y. Y.; Berger, R.; Holm, C.; Delle Site, L. *Faraday Discuss.* **2011**, DOI: 10.1039/C1FD00051A.

LISTb: a Better Direct Approach to LIST

Ya Kun Chen and Yan Alexander Wang*

Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, British Columbia, V6T 1Z1, Canada

ABSTRACT: Following our recent paper on linear-expansion shooting techniques (LIST) [Wang, Y. A.; Yam, C. Y.; Chen, Y. K.; Chen, G. H. *J. Chem. Phys.* 2011, 134, 241103], in which the direct approach (LISTd) and the indirect approach (LISTi) were proposed to accelerate the self-consistent field convergence, we discovered a highly simple solution to cure the linear-dependence problem of LISTd. The resultant method, LISTb, is a better direct approach to LIST and muscles similar performances to existing LIST methods. More promisingly, LISTb even outshines the best LIST method, LISTi, for systems involving transition-metal atoms.

The self-consistent field (SCF) method has been widely adopted to solve the Hartree–Fock and Kohn–Sham density-functional theory problems.¹ In the Roothaan–Hall procedure,¹ a usual SCF step involves rendering the input density matrix (D^{in}) into the Fock Hamiltonian matrix, then diagonalizing the Fock matrix to obtain the output density matrix (D^{out}) and generating an initial guess (input) for the next iteration. Writing this into equations,¹ for iteration i and orbital k , one has

$$\left\{ -\frac{1}{2}\nabla^2 + \hat{v}_{\text{eff}}[D^{\text{in}}] \right\} |\psi_{k,i}^{\text{out}}\rangle = \varepsilon_{k,i}^{\text{out}} |\psi_{k,i}^{\text{out}}\rangle \quad (1)$$

and

$$\sum_k^{\text{occ}} f_{k,i} |\psi_{k,i}^{\text{out}}\rangle \langle \psi_{k,i}^{\text{out}}| = D_i^{\text{out}} \quad (2)$$

where \hat{v}_{eff} is the effective potential, $f_{k,i}$ and $\varepsilon_{k,i}$ are the occupation number and the eigenvalue of orbital $\psi_{k,i}$ respectively, and the superscripts “out” and “in” stand for the output and input quantities, respectively. At iteration i , the electronic energy is normally evaluated via the Hohnberg–Kohn–Sham (HKS) energy functional:²

$$E_i^{\text{HKS}}[D_i^{\text{out}}] = \sum_k^{\text{occ}} f_{k,i}^{\text{out}} \varepsilon_{k,i}^{\text{out}} - \langle \hat{v}_{\text{eff}}[D_i^{\text{in}}] D_i^{\text{out}} \rangle + E_{\text{H}}[D_i^{\text{out}}] + E_{\text{XC}}[D_i^{\text{out}}] \quad (3)$$

where E_{H} and E_{XC} are the Hartree and exchange-correlation energies, respectively. However, such a pristine SCF algorithm only works for simple small species and fails in most contemporary quantum chemistry studies. There have been some great efforts in designing more powerful schemes to accelerate the convergence of the SCF process prior to 2011.^{3–13} Logically, such SCF acceleration schemes can be divided into two categories: ones that change the SCF route^{3–8} and ones that do not.^{9–13}

In the first category that involves changing the variational path, Pulay’s direct inversion in the iterative subspace (DIIS) algorithm is the most successful one and is widely used today.^{3,4} However, the standard DIIS algorithm does fail occasionally because the targeted commutativity between the Fock matrix and the density matrix is only a necessary but not a sufficient condition for SCF convergence toward an energy minimum. Alternatively, DIIS can be reformulated as a Krylov subspace accelerated inexact Newton (KAIN) method,⁵

but with limited success in performance improvement.⁶ To incorporate the condition of energy minimization, energy-DIIS (EDIIS) and augmented-Roothaan–Hall energy-DIIS (ADIIS) methods have been developed and have shown their ability in some challenging cases.^{7,8} Nonetheless, their strengths are more prominent in the early stages of SCF processes, whereas DIIS takes over the task of acceleration as the SCF is near the final convergence.⁸

The other technique that accelerates the total energy convergence is to only finesse the energy evaluation formula, which has been systematically investigated in our group.^{9–13} On the basis of the analysis of the difference between the final converged exact energy and the current-iteration HKS energy, the corrected HKS (cHKS) energy functional can be used to improve the evaluation of the total electronic energy.^{9,10}

$$E_i^{\text{cHKS}} = E_i^{\text{HKS}}[D_i^{\text{out}}] + \langle \Delta v_i (D^{\text{b}} - D_i^{\text{out}}) \rangle \quad (4)$$

with $\Delta v_i = (\hat{v}_{\text{eff}}[D_i^{\text{out}}] - \hat{v}_{\text{eff}}[D_i^{\text{in}}])/2$, where D^{b} denotes the best estimate of the final converged exact density matrix based on information currently available (before the convergence is reached).

Recently, two linear-expansion shooting techniques (LIST), namely, the direct approach (LISTd) and the indirect approach (LISTi), have been developed in our group.¹⁴ Within LISTd, we imposed the cHKS energy for any iteration to be equal to the final converged energy (exact up to second-order corrections), given that D^{b} is optimally expressed as a linear expansion of historical output density matrices. The final matrix equation for the LISTd method can be succinctly written as

$$\underbrace{\begin{pmatrix} 0 & -1 & -1 & \cdots & -1 \\ -1 & a_{11} & a_{12} & \cdots & a_{1m} \\ -1 & a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}}_{\text{A}} \underbrace{\begin{pmatrix} E \\ c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix}}_{\text{C}} = \underbrace{\begin{pmatrix} -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\text{O}}, \quad (5)$$

with $a_{ij} = E_i^{\text{HKS}}[D_i^{\text{out}}] + \langle \Delta v_i (D_j^{\text{out}} - D_i^{\text{out}}) \rangle$, where E is the current best estimate of the total energy and D_j^{out} is a historical

Received: June 29, 2011

Published: September 14, 2011

output density matrix used to expand D^b via the expansion coefficients $\{c_j\}$:

$$D^b = \sum_{j=1}^m c_j D_j^{\text{out}} \quad (6)$$

with the normalization condition

$$\sum_{j=1}^m c_j = 1 \quad (7)$$

In eq 5, one should note that matrix **A** for LISTd is asymmetric, which will become vitally important later.

Taking an alternative approximation, within LISTi, we imposed the equalization of the CHKS energies expanded in terms of the input and output density matrices, resulting in an almost identical matrix equation except that the matrix elements of **A** are defined as $\langle \Delta v_i (D_j^{\text{out}} - D_j^{\text{in}}) \rangle$. Solving the linear equations for LISTd or LISTi, one has a set of expansion coefficients $\{c_j\}$ to construct the input effective potential for the next iteration

$$\hat{v}_{\text{eff}}[D_{m+1}^{\text{in}}] = \sum_{j=1}^m c_j \hat{v}_{\text{eff}}[D_j^{\text{out}}] \quad (8)$$

Case studies have shown that LISTi is more effective than LISTd, because LISTd suffers from linear-dependence problems near convergence.¹⁴ Hereafter, we introduce a *better* direct approach to the LIST method, LISTb, which is capable of curing the linear-dependence problem rooted in LISTd.

We first start with the key imposition within LISTd:¹⁴

$$\sum_{j=1}^m c_j a_{ij} = E, \quad \forall i \quad (9)$$

Upon introducing an additional set of expansion coefficients $\{c'_i\}$ with $\sum_{i=1}^m c'_i = 1$, we can generalize eq 9 into a double summation form:

$$\sum_{i=1}^m \sum_{j=1}^m c'_i c_j a_{ij} = E \quad (10)$$

Now, interchanging the order of summations on the left-hand side of eq 10, we arrive at

$$\sum_{j=1}^m c_j \left\{ \sum_{i=1}^m c'_i a_{ij} \right\} = E \quad (11)$$

in which we can choose the values of $\{c'_i\}$ properly, such that

$$\sum_{i=1}^m c'_i a_{ij} = E', \quad \forall j \quad (12)$$

After plugging eq 12 back into eq 11, we readily conclude $E = E'$ because of eq 7. Then, eqs 9–12 can be recast in terms of a single expansion over $\{c'_i\}$ instead:

$$\sum_{i=1}^m c'_i a_{ij} = E, \quad \forall j \quad (13)$$

which immediately leads to a matrix equation virtually identical to eq 5:

$$\mathbf{B}\mathbf{C}' = \mathbf{A}^T \mathbf{C}' = \mathbf{O} \quad (14)$$

except that matrix **B** is exactly the transpose of matrix **A** ($b_{ij} = a_{ji}$) and $\mathbf{C}' = (E, c'_1, c'_2, \dots, c'_m)^T$. Of course, this is only meaningful if

matrix **B** differs from matrix **A**, which fortunately is so, as already noted above.

Amazingly, it turns out that this transposition is the key to success here: it avoids the severe linear dependence of LISTd while inheriting the power of LIST. The linear-dependence problem of LISTd stems from the fact that as the calculation approaches the final convergence, the rightmost column of matrix **A** becomes increasingly similar to the column on its left, with an element-wise difference

$$a_{i,m} - a_{i,m-1} = \langle \Delta v_i (D_m^{\text{out}} - D_{m-1}^{\text{out}}) \rangle \quad (15)$$

which is of second order of the density matrix residual error. On the other hand, LISTb will not suffer from the same problem of LISTd, because the difference between two neighboring columns of matrix **B**

$$b_{i,j} - b_{i,j-1} = (E_j^{\text{HKS}} - E_{j-1}^{\text{HKS}}) + \langle (\Delta v_j - \Delta v_{j-1}) D_i^{\text{out}} \rangle - \langle \Delta v_j D_j^{\text{out}} \rangle + \langle \Delta v_{j-1} D_{j-1}^{\text{out}} \rangle \quad (16)$$

is dominated by terms of first order of the density matrix residual error.

According to Cramer's rule,¹⁵ the solution to eq 5 or eq 14, $\{c_i\}$ for LISTd or $\{c'_i\}$ for LISTb, is the quotient of two determinants, with the numerator being the determinant of a matrix with one column replaced by vector **O** and the denominator being the determinant of matrix **A**. For example, $c_m = |\mathbf{A}_m|/|\mathbf{A}|$, where matrix \mathbf{A}_m is simply matrix **A** with its last column replaced by vector **O**. Therefore, in the LISTd scheme, as it approaches the final convergence, the tiny difference between the last two columns, eq 15, in matrix **A** will routinely result in c_{m-1} and c_m of very big, nearly identical magnitude but with opposite signs. In contrast, LISTb has a much bigger difference in the last two columns, eq 16, that avoids the build-up of the linear-dependence problem. Thus, the transpose of matrix **A** of LISTd into matrix **B** of LISTb yields a different, much improved acceleration path. For future reference, we name the above transpose technique "alleviation of linear dependence in asymmetric system via transposition" (ALDAST).

To compare LISTb with other LIST members and DIIS, all LIST methods, including LISTb, LISTd, and LISTi, were implemented in the NWChem 5.0 source code.¹⁶ Within a LIST calculation, the CHKS total electronic energy functionals were evaluated at every iteration.^{9,10} In all calculations, only five Fock and density matrices were used in the linear mixing unless otherwise noted, and no other SCF convergence schemes, such as level shifting or density damping, were invoked. For all calculations, full convergence was defined as the energy difference between two consecutive iterations smaller than 10^{-9} Hartrees.

We chose several molecules to benchmark the performance of our methods. For hydrogen fluoride (HF), the H–F bond length was 0.920 Å. For water (H₂O), the two H–O bond lengths and the H–O–H bond angle were 0.965 Å and 103.75°, respectively. For ethene (C₂H₄), the C–C bond length and all C–H bond lengths were 1.335 and 1.098 Å, respectively, while all H–C–C bond angles were 122.88°. The C–C and C–H bond lengths in benzene (C₆H₆) were all set to 1.396 and 1.097 Å, respectively, and all bond angles were 120°. For the aforementioned four molecules, a local density approximation (LDA) exchange-correlation functional was used in conjunction with the 6-31G basis set.^{17,18} A large water cluster system, (H₂O)₅₁, was also

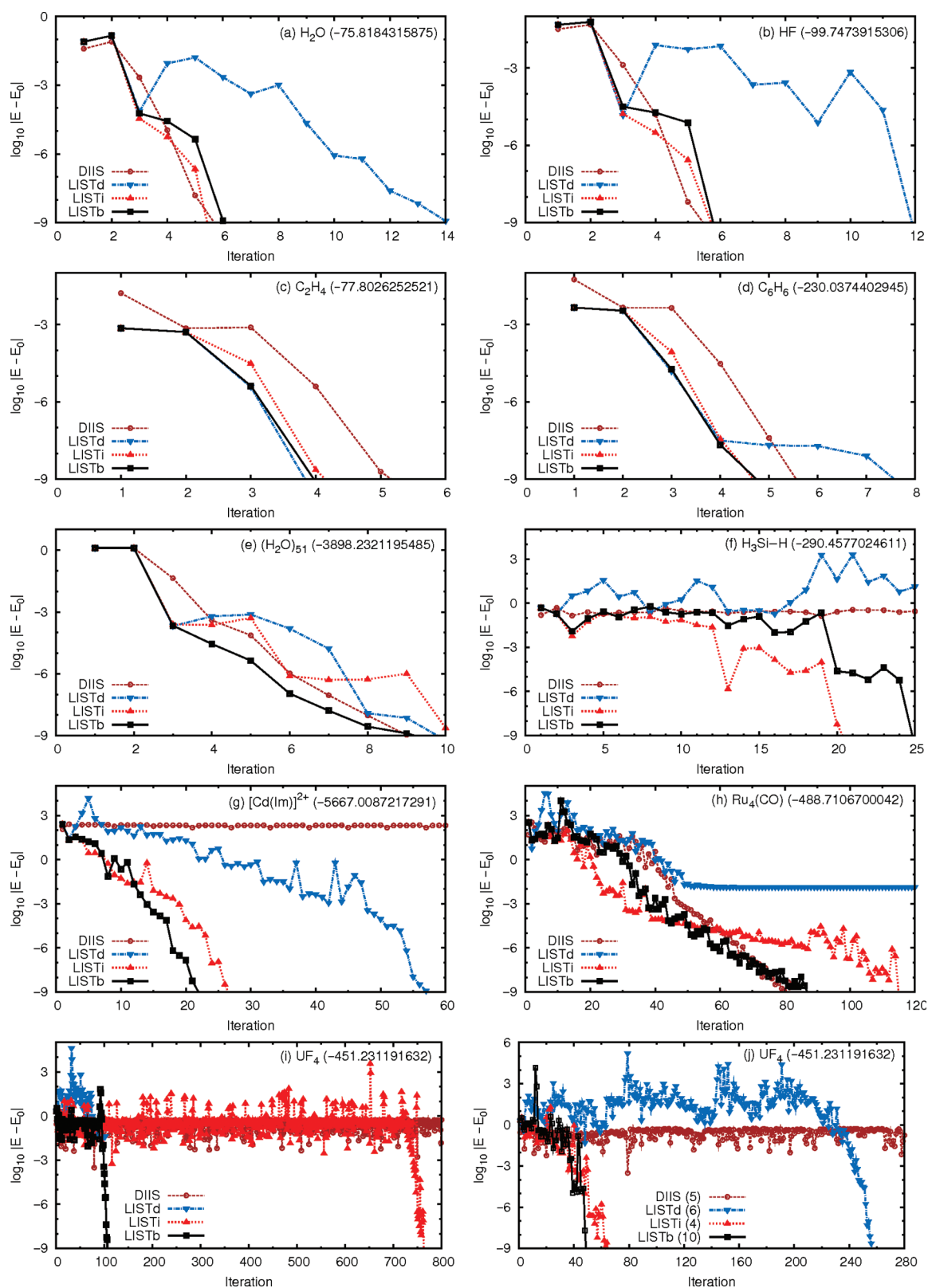


Figure 1. Convergence of the total energy (in Hartrees) for different systems. The final converged energy E_0 (in Hartrees) is in the parentheses in the topright corner of each subfigure. In j, the optimal linear-expansion length of each method is in the parentheses of the legend. In a–i, the linear-expansion length is fixed to 5.

calculated at the B3LYP/6-31G** level of theory,¹⁹ and its geometry can also be found in ref 8.

We have also performed numerical tests on several challenging systems that had been studied elsewhere under similar conditions.^{7–9,14}

A silane (SiH_4) molecule with a much elongated Si–H bond was calculated at the LDA/6-31G* level of theory, under the same conditions as previously studied:^{9,14} the three regular and the elongated Si–H bond lengths were 1.47 and 4.00 Å, respectively, and all four H–Si–H bond angles were 109.28°. A cadmium-imidazole cation ($[\text{Cd}(\text{Im})]^{2+}$)^{8,14} was calculated at the B3LYP/3-21G level of theory¹⁹ using the core Hamiltonian as the very first initial guess. The tetranuclear ruthenium carbonyl cluster $\text{Ru}_4(\text{CO})^8$ and tetrahedral uranium fluoride (UF_4) with a U–F bond length of 1.98 Å⁷ were also calculated at the B3LYP/LanL2DZ level with the core electrons of the transition-metal atoms represented by the LanL2 pseudopotentials.²⁰ For UF_4 , a superposition of atomic densities was utilized as the very first initial guess, different from ref 7.

The performances of different LIST methods and DIIS are compared in Figure 1. For simple molecules around their equilibrium geometries, such as HF and H_2O , LISTb, LISTi, and DIIS demonstrate comparable performances (maybe with a slightly slower pace for LISTb), and all three outpace LISTd by about eight iterations. Cases with more complex π bonding interactions, i.e., C_2H_4 and C_6H_6 , also confirm that LISTb is comparable to LISTd, LISTi, and DIIS: all of them can achieve a similar rate of convergence for SCF calculations, whereas LISTd is only marginally slower than the other three algorithms. For the much bigger system, $(\text{H}_2\text{O})_{51}$, LISTb unequivocally outperforms all other methods, whereas LISTi goes sideways after iteration six but catches up after iteration nine. The silane molecule with a much elongated Si–H bond represents a prototype case of nonequilibrium structures. Both DIIS and LISTd fail in this case, while LISTb and LISTi can converge within 25 iterations with LISTi leading by four iterations.

For molecules containing transition-metal atoms that involve much more complicated bonding interactions, LISTb is the clear winner among all four methods compared in this study. Despite the complete failure of DIIS for $[\text{Cd}(\text{Im})]^{2+}$, all three LIST methods work but at different speeds: LISTb can reach full convergence within only 22 iterations, five and 35 iterations faster than LISTi and LISTd, respectively. In the case of $\text{Ru}_4(\text{CO})$, all four methods converge, but LISTd guides the system toward a different energy of -488.6972714297 Hartrees, higher than the true ground state by about 0.0134 Hartrees. LISTi falls behind eventually even after outperforming others in the first 40 iterations. LISTb begins to lead the race after 50 iterations, while DIIS only catches up after nearly 80 iterations. As for UF_4 , LISTb only takes about 100 iterations to converge, whereas LISTi spends more than 760 iterations and DIIS does not converge for the first 1000 iterations. Again, the LISTd scheme in this case converges to a higher energy of -451.1890747131 Hartrees within 100 iterations.

We also found that the length of the linear expansion of eq 8 does affect the convergence behavior of all methods studied here. Taking $\text{Ru}_4(\text{CO})$ for example, DIIS only converges when the linear-expansion length is 5. For this very reason, we have intentionally presented the results with five expansion Fock matrices in Figure 1a–i. If we choose the fastest convergence rate of each method with its optimal expansion length, Figure 1j shows that, for UF_4 , LISTb and LISTi can reach full convergence within 49 and 65 iterations with 10 and 4 expansion Fock matrices, respectively. Interestingly, LISTd does converge with six expansion Fock matrices, but only after 256 iterations. It is thus desirable to finetune the length of the linear expansion to achieve optimal performance, although five expansion Fock matrices seem to be a good default choice for most systems investigated thus far.

Overall, the above case studies indicate that LISTb rivals LISTi in terms of the robustness and effectiveness of accelerating SCF convergence, especially for species containing transition-metal atoms. Though LISTb shares the same overall matrix equation with LISTd and LISTi, but with the ALDAST technique curing the severe linear-dependence problem faced by LISTd, LISTb delivers a more smooth, much faster SCF convergence path, particularly illustrated by the $\text{Ru}_4(\text{CO})$ and UF_4 calculations. On the basis of the numerical tests reported in this work and previous publication,¹⁴ we now wholeheartedly endorse LISTb and LISTi as the two most powerful implementations of LIST.

AUTHOR INFORMATION

Corresponding Author

*E-mail: yawang@chem.ubc.ca.

ACKNOWLEDGMENT

Financial support for this study was provided by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

- (1) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; Wiley: West Sussex, U.K., 2007; pp 80–115.
- (2) Chelikowsky, J. R.; Louie, S. G. *Phys. Rev. B* **1984**, *29*, 3470–3481.
- (3) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393–398.
- (4) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556–560.
- (5) Harrison, R. J. *J. Comput. Chem.* **2004**, *25*, 328–334.
- (6) Rohwedder, T.; Schneider, R. *J. Math. Chem.* **2011**, *49*, 1889–1914.
- (7) Kudin, K. N.; Scuseria, G. E.; Cancés, E. *J. Chem. Phys.* **2002**, *116*, 8255–8261.
- (8) Hu, X.; Yang, W. *J. Chem. Phys.* **2010**, *132*, 054109.
- (9) Zhang, Y. A.; Wang, Y. A. *J. Chem. Phys.* **2009**, *130*, 144116.
- (10) Zhou, B.; Wang, Y. A. *J. Chem. Phys.* **2008**, *128*, 084101.
- (11) Zhou, B.; Wang, Y. A. *J. Chem. Phys.* **2007**, *127*, 064101.
- (12) Zhou, B.; Wang, Y. A. *J. Chem. Phys.* **2006**, *124*, 081107.
- (13) Zhou, B.; Wang, Y. A. *Int. J. Quantum Chem.* **2007**, *107*, 2995–3000.
- (14) Wang, Y. A.; Yam, C. Y.; Chen, Y. K.; Chen, G. H. *J. Chem. Phys.* **2011**, *134*, 241103.
- (15) Weber, H. J.; Arfken, G. B. *Essential Mathematical Methods for Physicists*; Academic Press: San Diego, CA, 2003; pp 162–168.
- (16) Bylaska, E. J.; de Jong, W. A.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, E.; Windus, T. L.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Frühtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem 5.0*; Pacific Northwest National Laboratory: Richland, Washington, 2006.
- (17) Slater, J. C.; Johnson, K. H. *Phys. Rev. B* **1972**, *5*, 844–853.
- (18) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1210.
- (19) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5672.
- (20) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.

Perferryl Fe^V–Oxo Nonheme Complexes: Do They Have High-Spin or Low-Spin Ground States?

Hui Chen,^{*,†,‡} Wenzhen Lai,[†] Jiannian Yao,[‡] and Sason Shaik^{*,†}

[†]Institute of Chemistry and the Lise Meitner-Minerva Center for Computational Quantum Chemistry, Hebrew University of Jerusalem, Givat Ram Campus, 91904 Jerusalem, Israel

[‡]Beijing National Laboratory for Molecular Sciences (BNLMS), CAS Key Laboratory of Photochemistry, Institute of Chemistry, Chinese Academy of Sciences, Beijing, 100190, China

S Supporting Information

ABSTRACT: Nonheme perferryl Fe^V=O species are studied herein by means of coupled cluster (CCSD(T)) calculations with a complete basis set limit estimate and density functional B3LYP computations. It is shown that the high-spin/low-spin (HS/LS) energy order in these Fe^V=O species is highly dependent on the electronic nature of the ligand sphere and the geometric position of ligands relative to the Fe^V=O moiety. When only σ -donor amines ligate Fe^V=O, the LS state is slightly lower than the HS states. However, when a strong π -donor ligand such as hydroxyl is cis to Fe^V=O, the HS state becomes highly favored. And on the contrary, if the π -donor ligand is trans to Fe^V=O, the LS state is predicted here to be highly favored. This last type of perferryl complex has not yet been made by experimental means. Generally, our findings are consistent with the available experimental data.^{4a,6,7} Some implications of these findings on the behavior of experimental systems are discussed.

Nonheme high-valent oxoiron reagents have attracted considerable interests in bioinorganic chemistry, since these complexes mimic active oxidants of many nonheme enzymes.¹ Among the high-valent oxoiron complexes, the ferryl species have been extensively studied both experimentally² and theoretically.³ These studies have generated a great deal of knowledge about the spin-state energies and oxidative reactivity of ferryl Fe^{IV}=O species. By contrast, our knowledge of perferryl Fe^V=O species remains scant. The formation of perferryl Fe^V=O was proposed, for some nonheme enzymes and synthetic analogues, based on indirect experimental indications.⁴ To date, the only experimentally observed perferryl species is the synthetic complex [O=Fe^V(TAML)]⁻ of Collins et al.,⁵ wherein TAML is a macrocyclic tetraamide ligand of the tetra anion (see Chart 3). However, there are indications that more perferryl species might have been observed, but their structures and electronic features are still not firmly established. Hence, a study of their types and their spin states is deemed timely.

Very recently, there appeared an experimental EPR study⁶ which assigned the spin-state identity of the proposed⁷ reactive Fe^V=O species during nonheme iron catalyzed epoxidation of alkenes. Thus, Talsi et al.⁶ showed that active species of the catalyst systems based on nonheme iron complexes, tentatively assignable to be Fe^V=O, are low-spin species. However, it remains unclear whether and how the spin state ordering of perferryl Fe^V=O is determined by the ligand sphere. In response, we present a high level ab initio coupled cluster (CC) with complete basis set (CBS) limit estimate as well as DFT computational results for electronic structures and spin state energetics of nonheme perferryl Fe^V=O complexes.⁸ The study accounts for and rationalizes the experimental findings for the related experimental species and uncovers new features. As we shall demonstrate,

the perferryls turn out to possess a spectrum of intriguing features. The usage of CC theory further allows us to evaluate the performance of an approximate functional like B3LYP in describing all of the available electronic states of perferryls. The computational methods and details (structures, energies, tests of the CBS limit estimates, potential energy surface scan profile) are all given in the Supporting Information (SI).

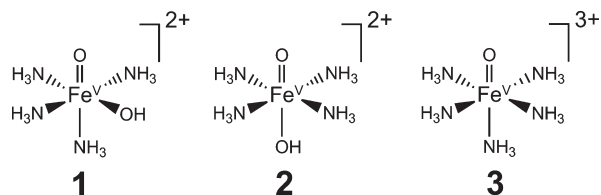
A prototype perferryl Fe^V=O structure that has been proposed frequently in experimental systems^{4,6,7} is [O=Fe^V-OH(L)]²⁺, wherein the hydroxyl group is at the cis position to the Fe=O unit, and L is the remaining neutral polydentate ligand made usually of amine/pyridine moieties. Alternatively, there could also exist perferryl complexes devoid of an OH ligand, i.e., [O=Fe^V(L)]³⁺, wherein only a neutral polydentate ligand, or maybe one additional solvent molecule, is wrapped around Fe^V=O. Moreover, changing the cis-OH to trans-OH will generate a third complex, with electronic structure features that are not known from experiments. As such, we designed the three hexa-coordinate model complexes shown in Chart 1, for which high-level CCSD(T) calculations with a CBS limit extrapolation⁹ estimate can be applied to serve as our reference for DFT calculations.

The various electronic configurations of perferryl complexes are depicted in Scheme 1. Generally, the high-spin (HS) quartet (Q) states (Q_{xy}, Q_{pO}, and Q_{xy-pO}) are all characterized by triradicaloid configurations, while the LS doublet (D) states can be classified as either a monoradicaloid configuration (D _{π}) with one unpaired electron or triradicaloid doublet configurations (D_{pO} and D_{xy-pO}) with three unpaired electrons coupled

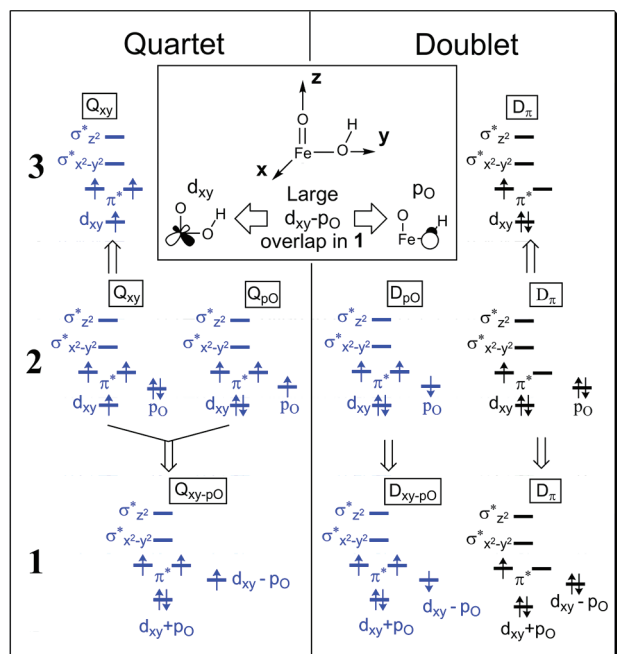
Received: August 31, 2011

Published: September 09, 2011

Chart 1. The Three Used Model Perferryl Complexes



Scheme 1. Potential Triradicaloid (blue) and Monoradicaloid (black) Configurations for 1–3 and Their Connections



antiferromagnetically. Previous DFT calculations focused mostly on the triradicaloid LS states.^{10a} One exception is the study by Neese et al. for **2**, wherein the monoradicaloid LS state was calculated while the triradicaloid LS state was not reported.¹¹

The configurations in Scheme 1 are system-dependent and fall into two classes: One class is labeled Q_{pO} , Q_{xy-pO} , D_{pO} , and D_{xy-pO} , wherein the subscripts indicate that OH acts as a noninnocent ligand. This arises because the short Fe–OH bond (1.8 Å) creates a large overlap between the $p(OH)$ orbital, perpendicular to the Fe–O–H plane, and the π^* or d_{xy} type Fe=O orbital, whereby the OH ligand acquires also significant spin density. The other class is labeled as Q_{xy} , D_{π} , which means that the open shell part of these configurations involves only the orbitals of the Fe^V=O moiety (d_{xy} or π^* FeO type, respectively). This latter class is also called “genuine” or sometimes “true” Fe^VO,¹² to distinguish it from the cases with noninnocent ligands, e.g., in the extreme case of compound **1** (Cpd I) in heme enzymes like P450, which possess a ferryl Fe^{IV}O unit and a porphyrin cation radical.¹³ Due to the absence of the π -donor hydroxyl ligand in **3**, the space separation of the unpaired electrons by ligand noninnocence is not possible anymore, and hence, the triradicaloid states must have their three unpaired electrons localized on the Fe^VO center. As such, the LS

Table 1. The Calculated Adiabatic Q–D Gaps (kcal/mol) for 1–3^a

complex	state	UB3LYP/def2-QZVP ^b	CCSD(T)/CBS
1	Q_{xy-pO}	0.0 (0.0) [0.0]	0.0
	D_{π}	10.9 (15.8) [9.7]	12.4
	D_{xy-pO}	16.3 (10.9) [15.9]	— ^c
2	Q_{xy}	0.0 (0.0) [0.0]	0.0
	Q_{pO}	−4.0 [−3.4]	1.5
	D_{π}^d	−4.3 [−3.4]	−15.8
	D_{pO}	−7.5 (−5.9) [−7.1]	— ^c
	D_{π}	−0.8 (4.7) [−0.8]	−3.1

^aThe structures used in CCSD(T) calculation are from UB3LYP/def2-TZVP geometry optimization for the corresponding states. ^bValues outside/inside parentheses are with/without spin-projection for spin contamination; values in brackets also include zero point energy (ZPE) correction for the spin-projected energies. ^cSymmetry-broken states for which the RCCSD(T) method cannot be applied. ^dNot spin contaminated.

triradicaloid states of **3** will be destabilized relative to the HS states by a loss of exchange energy.¹⁴ Consequently, in the absence of the π -donor OH ligand, the favored LS state in **3** will be the monoradicaloid LS state. Indeed, the only LS state which we could locate herein for **3** using B3LYP calculations is D_{π} .

Table 1 summarizes the calculated adiabatic Q–D gaps for **1–3** at the CCSD(T)/CBS estimate and the B3LYP/def2-QZVP levels. Inspection of the CCSD(T) results for **1**, for which the π -donor hydroxyl group is at the cis position to the perferryl Fe=O unit, shows that the ground state is HS and is preferred by 12.4 kcal/mol over the LS D_{π} state. By contrast to **1**, for **2**, wherein the π -donor hydroxyl ligand is trans to the perferryl Fe=O moiety, the LS D_{π} state is favored over the two HS states by 15.8 and 17.3 kcal/mol. Finally, for species **3**, in which only σ -donor amine ligands bind iron, and which may be considered as a model for nitrogenous-ligated nonheme complexes,¹ the HS and LS states are close in energy, with the latter being 3.1 kcal/mol lower. We note that this result for **3** is quite similar to our recent multireference CASPT2^{15a} computational results for the “true” Fe^VO electromers of Cpd I in heme systems, which has been subsequently confirmed also by RASPT2^{15b} computations. This, and other tests of the basis sets (see SI), show that the CCSD(T)/CBS estimates are solid.

For **1** and **3**, in Table 1, B3LYP performs in agreement with CCSD(T), especially after correction of the spin-contamination.¹⁶ Thus, complex **1** exhibits in spin-contamination corrected-B3LYP a monoradicaloid D_{π} state, which is 10.9 kcal/mol higher than the triradicaloid HS ground state,¹⁷ in reasonable agreement with the CCSD(T) datum of 12.4 kcal/mol. Similarly, for **3**, spin-contamination corrected B3LYP predicts a D_{π} ground state in agreement with CCSD(T). For **2**, the B3LYP predicts correctly the preference for a LS state. But the B3LYP energy gap relative to the HS state Q_{xy} is significantly smaller than the corresponding CCSD(T) gap. That this gap is large is supported by independent multireference SORCI calculation of the vertical gap (12.9 kcal/mol).^{11,18} This indicates that the performance of an approximate functional tends to be system-dependent. In addition, we note that the ZPE correction in Table 1 does not change our results much.

Since **1** and **2** are structural isomers, their relative energy is chemically relevant. Figure 1 shows this comparison, using as a

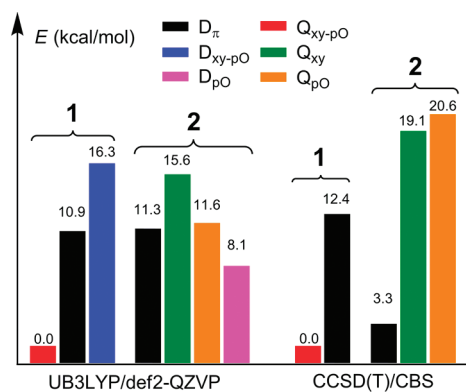
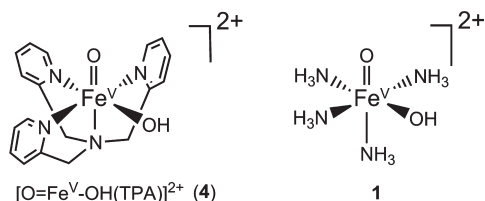


Figure 1. Relative UB3LYP/def2-QZVP and CCSD(T)/CBS estimate energies of various spin states for the isomeric perferryls, **1** and **2**, relative to the lowest state Q_{xy-pO} of **1**. Whenever necessary, the relative DFT energies of doublet states given are after spin-contamination correction.

Chart 2. Complex $[O=Fe^V-OH(TPA)]^{2+}$ (4**) and Its Model, **1****



reference the most stable state, Q_{xy-pO} for **1**, and depicts the relative energies of the other states. The computed stability of $1-Q_{xy-pO}$ is in agreement with the fact that all of the so far proposed nonheme $HO-Fe^V=O$ complexes under experimental conditions contain a cis hydroxyl ligand but never a trans OH ligand. Additionally, as shown in Figure 1, the difference between B3LYP and CCSD(T) for **2** is quite large. This large difference owes its origins to the following factors: one is the underestimation by B3LYP of the stability of $2-LS$; the other is overestimation of the stability of $2-HS$ states. With CCSD(T), the $2-D_{\pi}$ is only 3.3 kcal/mol higher than $1-Q_{xy-pO}$. As such, **2** may be thermodynamically accessible, although in small quantities. Furthermore, the conversion of $1-Q_{xy-pO}$ to $2-D_{\pi}$ may represent a new spin crossover system but for a high-valent iron, wherein the spin state interchanges as the OH ligand flips between the cis and trans positions.

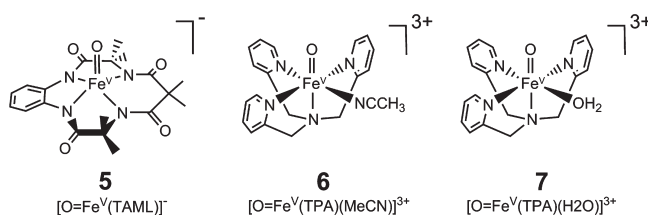
Having shown the key findings of the three perferryl model systems, one may wonder how relevant these findings are to the more realistic synthetic nonheme complexes. To answer this concern, we compare the predictions for **1** to a related experimentally proposed^{4a,10a–10c} complex, $[O=Fe^V-OH(TPA)]^{2+}$ (TPA = tris(2-pyridylmethyl)amine), **4**, shown in Chart 2. Table 2 shows the relative spin-state energies for **1** and **4**.¹⁹ The computed triradicaloid $Q_{xy-pO}-D_{xy-pO}$ gap for **4** (18.4 kcal/mol) is close to the previous B3LYP calculation (20.5 kcal/mol) using a smaller basis set,^{10a} and $Q_{xy-pO}-D_{\pi}$ gaps of both **1** and **4** in Table 2 are on par with the CCSD(T)/CBS gap for **1** in Table 1. Furthermore, it is apparent from Table 2 that the spin-contamination corrected gaps, $Q_{xy-pO}-D_{xy-pO}$ and $Q_{xy-pO}-D_{\pi}$, for **1** and **4** are very similar. As such, we may conclude that the simplified complex **1** is a

Table 2. The B3LYP Calculated Adiabatic Q–D Gaps (kcal/mol) for **1 and **4**^a**

	state	4	1
UB3LYP/def2-TZVPP ^b	Q_{xy-pO}	0.0 (0.0) [0.0]	0.0 (0.0)
	D_{π}	16.0 (20.5) [16.3]	11.1 (16.0)
	D_{xy-pO}	18.4 (12.2) [19.1]	16.1 (10.7)

^aThe structures in all calculations are optimized by UB3LYP/def2-TZVP for the corresponding states. ^bValues outside/inside parentheses are with/without spin-projection for spin contamination; values in brackets are from the SMD solvation model calculation with spin-projection for spin contamination.

Chart 3. Perferryl Complexes **5–7**



fairly reliable model of **4** as far as the spin-state ordering is concerned. Furthermore, since the experiments with **4** were carried out^{4a} in acetonitrile, we tested these spin-state gaps for **4** using the SMD continuum solvation model.²⁰ The results, which are shown in brackets in Table 2, demonstrate that the Q–D gaps are intrinsic properties of **4** and are almost unaffected by solvation. Generally, our consistent DFT and CCSD(T) calculations for **4** and **1** imply that **4** is not likely to have a LS ground state, which does not support the recent tentative proposal from Talsi et al.²¹ Their assignment may refer to a different structure than **4**.

Having shown our results for **1–4** including both model and experimental systems, we may now compare our CCSD(T) results to experimental findings.^{5,6} Thus, our results for **3** (in Table 1) indicate that whenever the ligand sphere of the perferryl moiety involves only σ -donor amine ligands, the corresponding $Fe^V=O$ complex is likely to be LS, with a very close HS state of the Q_{xy} type. The first part of this prediction is consistent with the proven LS identity of $[O=Fe^V(TAML)]^-$ found by Collins et al. and is depicted in **5** in Chart 3, albeit having a ligand arrangement different from that of **3**.⁵ Furthermore, Talsi et al.⁶ assigned, based on EPR spectra, a LS ground state, for complexes of the $[O=Fe^V-(L)(S)]^{3+}$ type, wherein L is an amine/pyridine ligand and S is acetonitrile or a water molecule, as in **6** and **7**.²² However, the second part of the prediction regarding the accessibility of the HS state has not been addressed yet by experimentation. This HS state may well affect the EPR or Mössbauer spectra of these complexes,^{24,25} and furthermore, being so low in energy, it is likely that this state will be involved in the reactivity of these perferryl reagents.

Our result of **2** further suggests that if an –OH or –OR ligand can be placed trans to the $Fe^V=O$ moiety, then the LS state will be strongly preferred over the HS (Table 1). This prediction awaits future experimental tests.

In conclusion, using coupled cluster and DFT calculations, we demonstrated here that the spin-state energetics of perferryl $Fe^V=O$ species is highly dependent on the ligand sphere and its arrangement in space. When there are no π -donor ligands such as

a –OH or –OR group but only σ -donor amine/pyridine ligands, the HS and LS perferryl states are close in energy with a slight preference for LS. When a π -donor ligand is at the cis position of perferryl $\text{Fe}^{\text{V}}=\text{O}$, the HS ground state is highly preferred, and on the contrary, when π -donor ligands are at the trans position of perferryl $\text{Fe}^{\text{V}}=\text{O}$, a LS ground state is highly preferred. The solvent has a minor influence on these spin state energetics of perferryl species. These theoretical findings can be used as fingerprints for probing perferryl complexes, which may or may not be present in a given medium.^{6,21} At present, the predictions are consistent with the previous observations which propose perferryl $\text{Fe}^{\text{V}}=\text{O}$ as an active species in those nonheme systems.⁷ Furthermore, the different spin states of these complexes may result in different reactivity patterns and may be worthy of experimental exploration.

■ ASSOCIATED CONTENT

S Supporting Information. Cartesian coordinates, computational details, and full set of computational results. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chenh@iccas.ac.cn; sason@yfaat.ch.huji.ac.il.

■ ACKNOWLEDGMENT

This work was supported by the Chinese Academy of Sciences (to H.C.) and ISF Grant 53/09 (to S.S.).

■ REFERENCES

- (1) (a) Nam, W. *Acc. Chem. Res.* **2007**, *40*, 522–531. (b) Que, L., Jr. *Acc. Chem. Res.* **2007**, *40*, 493–500. (c) Shan, X.; Que, L., Jr. *J. Inorg. Biochem.* **2006**, *100*, 421–433. (d) Que, L., Jr.; Tolman, W. B. *Nature* **2008**, *455*, 333–340. (e) Kryatov, S. V.; Rybak-Akimova, E. V.; Schinder, S. *Chem. Rev.* **2005**, *105*, 2175–2226.
- (2) (a) Costas, M.; Mehn, M. P.; Jensen, M. P.; Que, L., Jr. *Chem. Rev.* **2004**, *104*, 939–986. (b) Kovaleva, E. G.; Neiberger, M. B.; Chakrabarty, S.; Lipscomb, J. D. *Acc. Chem. Res.* **2007**, *40*, 475–483. (c) Krebs, C.; Fujimori, D. G.; Walsh, C. T.; Bollinger, J. R. *Acc. Chem. Res.* **2007**, *40*, 484–492. (d) Brujininx, P. C. A.; van Koten, G.; Gebbink, R. J. M. K. *Chem. Soc. Rev.* **2008**, *37*, 2716–2744. (e) Ryle, M. J.; Hausinger, R. P. *Curr. Opin. Chem. Biol.* **2002**, *6*, 193–201. (f) Abu-Omar, M. M.; Loaiza, A.; Hontzeas, N. *Chem. Rev.* **2005**, *105*, 2227–2252. (g) Solomon, E. I.; Brunold, T. C.; Davis, M. I.; Kemsley, J. N.; Lee, S.-K.; Lehnert, N.; Neese, F.; Skulan, A. J.; Yang, Y.-S.; Zhou, J. *Chem. Rev.* **2000**, *100*, 235–349. (h) Solomon, E. I.; Wong, S. D.; Liu, L. V.; Decker, A.; Chow, M. S. *Curr. Opin. Chem. Biol.* **2009**, *13*, 99–113. (i) Borovik, A. S. *Chem. Soc. Rev.* **2010**, *40*, 1870–1874. (j) Pestovsky, O.; Bakac, A. *J. Am. Chem. Soc.* **2004**, *126*, 13757–13764.
- (3) (a) Shaik, S.; Hirao, H.; Kumar, D. *Acc. Chem. Res.* **2007**, *40*, 532–542. (b) Bassan, A.; Blomberg, M. R. A.; Borowski, T.; Siegbahn, P. E. M. *J. Inorg. Biochem.* **2006**, *100*, 727–743. (c) Schöneboom, J. C.; Neese, F.; Thiel, W. *J. Am. Chem. Soc.* **2005**, *127*, 5840–5853. (d) Decker, A.; Clay, M. D.; Solomon, E. I. *J. Inorg. Biochem.* **2006**, *100*, 697–706. (e) Decker, A.; Rohde, J.-U.; Klinker, E. J.; Wong, S. D.; Que, L., Jr.; Solomon, E. I. *J. Am. Chem. Soc.* **2007**, *129*, 15983–15996. (f) Decker, A.; Clay, M. D.; Solomon, E. I. *J. Inorg. Biochem.* **2006**, *100*, 697–706. (g) Ye, S.; Neese, F. *Curr. Opin. Chem. Biol.* **2009**, *13*, 89–98. (h) Banse, F.; Girerd, J.-J.; Robert, V. *Eur. J. Inorg. Chem.* **2008**, 4786–4791. (i) Ensing, B.; Buda, F.; Gribnau, M. C. M.; Baerends, E. J. *J. Am. Chem. Soc.* **2004**, *126*, 4355–4365.
- (4) (a) Das, P.; Que, L., Jr. *Inorg. Chem.* **2010**, *49*, 9479–9485 and references therein. (b) Costas, M.; Chen, K.; Que, L., Jr. *Coord. Chem. Rev.* **2000**, *200*, 517–544. (c) Yoon, J.; Wilson, S. A.; Jang, Y. K.; Seo, M. S.; Nehru, K.; Hedman, B.; Hodgson, K. O.; Bill, E.; Solomon, E. I.; Nam, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1257–1260. (d) Makhlynets, O. V.; Rybak-Akimova, E. V. *Chem.—Eur. J.* **2010**, *16*, 13995–14006. (e) Lee, S. H.; Han, J. H.; Kwak, H.; Lee, S. J.; Lee, E. Y.; Kim, H. J.; Lee, J. H.; Bae, C.; Lee, S. N.; Kim, Y.; Kim, C. *Chem.—Eur. J.* **2007**, *13*, 9393–9398. (f) Chow, T. W.-S.; Wong, E. L.-M.; Guo, Z.; Liu, Y.; Huang, J.-S.; Che, C.-M. *J. Am. Chem. Soc.* **2010**, *132*, 13229–13239. (g) Li, F.; Meier, K. K.; Cranswick, M. A.; Chakrabarti, M.; Van Heuvelen, K. M.; Münck, E.; Que, L., Jr. *J. Am. Chem. Soc.* **2011**, *133*, 7256–7259.
- (5) de Oliveira, F. T.; Chanda, A.; Banerjee, D.; Shan, X.; Mondal, S.; Que, L., Jr.; Bominaar, E. L.; Münck, E.; Collins, T. J. *Science* **2007**, *315*, 835–838.
- (6) Lyakin, O. Y.; Bryliakov, K. P.; Britovsek, G. J.; Talsi, E. P. *J. Am. Chem. Soc.* **2009**, *131*, 10798–10799.
- (7) Mas-Ballesté, R.; Que, L., Jr. *J. Am. Chem. Soc.* **2007**, *129*, 15964–15972.
- (8) For previous CC calculations for $\text{Fe}^{\text{IV}}\text{O}$ models, see: (a) Chen, H.; Lai, W. Z.; Shaik, S. *J. Phys. Chem. Lett.* **2010**, *1*, 1533–1540. (b) Geng, C. Y.; Ye, S. F.; Neese, F. *Angew. Chem., Int. Ed.* **2010**, *49*, 5717–5720.
- (9) See, for example: (a) Carreón-Macedo, J.-L.; Harvey, J. N. *Phys. Chem. Chem. Phys.* **2006**, *8*, 93–100. (b) Oláh, J.; Harvey, J. N. *J. Phys. Chem. A* **2009**, *113*, 7338–7345. (c) Strickland, N.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841–852.
- (10) (a) Bassan, A.; Blomberg, M. R. A.; Siegbahn, P. E. M.; Que, L., Jr. *Chem.—Eur. J.* **2005**, *11*, 692–705. (b) Bassan, A.; Blomberg, M. R. A.; Siegbahn, P. E. M.; Que, L., Jr. *J. Am. Chem. Soc.* **2002**, *124*, 11056–11063. (c) Bassan, A.; Blomberg, M. R. A.; Siegbahn, P. E. M.; Que, L., Jr. *Angew. Chem., Int. Ed.* **2005**, *44*, 2939–2941. (d) Quiñero, D.; Morokuma, K.; Musaev, D. G.; Mas-Ballesté, R.; Que, L., Jr. *J. Am. Chem. Soc.* **2005**, *127*, 6548–6549. (e) Ma, Y.; Balbuena, P. B. *J. Phys. Chem. B* **2007**, *111*, 2711–2718. (f) de Visser, S. P.; Nam, W. *J. Phys. Chem. A* **2008**, *112*, 12887–12895.
- (11) Berry, J. F.; DeBeer George, S.; Neese, F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4361–4374.
- (12) (a) Dey, A.; Ghosh, A. *J. Am. Chem. Soc.* **2002**, *124*, 3206–3207. (b) Wasbotten, I.; Ghosh, A. *Inorg. Chem.* **2006**, *45*, 4910–4913.
- (13) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 949–1017.
- (14) Shaik, S.; Chen, H.; Janardanan, D. *Nature Chem.* **2011**, *3*, 19–27.
- (15) (a) Chen, H.; Song, J. S.; Lai, W. Z.; Wu, W.; Shaik, S. *J. Chem. Theory Comput.* **2010**, *6*, 940–953. (b) Radóń, M.; Broclawik, E.; Pierloot, K. *J. Chem. Theory Comput.* **2011**, *7*, 898–908.
- (16) Due to the higher energy of D_{π} compared to D_{xy-pO} before spin-contamination correction, DFT calculation may meet with a convergence problem without utilizing molecular symmetry, and hence calculating D_{π} poses a technical challenge for DFT.
- (17) In contrast to D_{pO} and D_{xy-pO} , for which the LS triradicaloid (hence spin-contaminated) configurations are shown in Scheme 1, it is not obvious that D_{π} is also spin-contaminated. It is an interesting observation in DFT calculations for the D_{π} state of **1** and **3** (but not **2**) that symmetry breakage happens to a certain extent (see $\langle S^2 \rangle$ in Table S1 in Supporting Information). For D_{π} , this occurs in a way that, in one of the two perpendicular $\text{Fe}=\text{O}$ π/π^* orbital sets, with a total occupation number of 2.0, the π bonding-orbital becomes less bonding due to partial α and β orbital localization on Fe and O, respectively, which can be visualized in the following way: $\uparrow\text{Fe}(\text{V})=\text{O} \rightarrow \uparrow\text{Fe}(\text{V})\bullet-\bullet\text{O}$, as a transition to more covalency, see, Isobe, H.; Yamanaka, S.; Okumura, M.; Yamaguchi, K.; Shimada, J. *J. Phys. Chem. B* **2011**, *115*, 10730–10738. Using natural orbitals, this symmetry breakage shows itself by having partial occupancy in π^* , which corresponds to some

multireference character of this $\pi(\text{FeO})$ bond, as previously analyzed for $\text{Fe}-\text{O}_2$ bonding in iron–oxy complexes in: Chen, H.; Ikeda-Saito, M.; Shaik, S. *J. Am. Chem. Soc.* **2008**, *130*, 14778–14790. Nevertheless, spin contamination caused by this symmetry breakage does not affect the important geometric parameters of the D_{3h} state, such as the $\text{Fe}-\text{O}$ bond distance (see Figure S1 in the Supporting Information).

(18) Note that the SORCI adiabatic gap for **2** (not reported in ref 11) will be larger than the vertical gap based on the HS geometry reported therein.

(19) Note that a smaller basis set is used in Table 2 than in Table 1 for DFT calculations, which can serve as an indication of near convergence for the basis set in our DFT calculations.

(20) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(21) Lyakin, O. Y.; Bryliakov, K. P.; Talsi, E. P. *Inorg. Chem.* **2011**, *50*, 5526–5538.

(22) We have tried to calculate by means of UB3LYP complexes **6** and **7**, which have higher positive charges ($3+$) than **4**. However, we found that in the case of the LS state, unpaired electrons undergo delocalization into the TPA ligand (pyridine ring), possibly due to the known self-interaction error (SIE) of the approximate functional, as found before in the hydrogen abstraction of weak C–H bonds for positively charged Fe^{IV} –oxo species.²³ Even in implicit solvent model calculations, the delocalization persists. This result cannot be considered reliable unless the SIE suspicion can be removed, which is not a simple matter. Thus, accurate DFT calculation of **6** and **7** may remain very challenging.

(23) (a) Johansson, A. J.; Blomberg, M. R.; Siegbahn, P. E. M. *J. Phys. Chem. C* **2007**, *111*, 12397–12406. (b) Johansson, A. J.; Blomberg, M. R.; Siegbahn, P. E. M. *J. Chem. Phys.* **2008**, *129*, 154301.

(24) Neese, F. *Coord. Chem. Rev.* **2009**, *253*, 526–563.

(25) (a) Rohde, J.-U.; In, J.-H.; Lim, M. H.; Brennessel, W. W.; Bukowski, M. R.; Stubna, A.; Münck, E.; Nam, W.; Que, L., Jr. *Science* **2003**, *299*, 1037–1039. (b) Bukowski, M. R.; Keohntop, K. D.; Stubna, A.; Bominaar, E. L.; Halfen, J. A.; Münck, E.; Nam, W.; Que, L., Jr. *Science* **2005**, *310*, 1000–1002. (c) Pestovsky, O.; Stoian, S.; Bominaar, E. L.; Shan, X.; Münck, E.; Que, L., Jr.; Bakac, A. *Angew. Chem., Int. Ed.* **2005**, *44*, 6871–6874.

Structural and Vibrational Properties of Liquid Water from van der Waals Density Functionals

Cui Zhang,[†] Jun Wu,[‡] Giulia Galli,^{*,†,§} and François Gygi^{‡,||}

[†]Department of Chemistry, University of California, Davis, California 95616, United States

[‡]Department of Applied Science, University of California, Davis, California 95616, United States

[§]Department of Physics, University of California, Davis, California 95616, United States

^{||}Department of Computer Science, University of California, Davis, California 95616, United States

S Supporting Information

ABSTRACT: We present results for the structural and vibrational properties of the water molecule, water dimer, and liquid water at the experimental equilibrium density, as obtained with several van der Waals density functionals. The functional form originally proposed by Dion et al. [*Phys. Rev. Lett.* **2004**, *92*, 246401], with an appropriately chosen local exchange functional, yields a description of the liquid superior to that of the semilocal functional PBE. In particular, a specific choice of the local exchange functional (optB88) fitted to quantum chemistry calculations yields the best agreement with experimental results for pair correlation functions although it is slightly inferior to other van der Waals functionals in describing infrared spectra. When using optB88, liquid water displays a hydrogen-bonded network less tightly bound than when using the PBE approximation. The performance of optB88 is definitely inferior to that of the PBE0 hybrid functional for the isolated molecule but only moderately so for the liquid. However, the computational cost of optB88 is much less than that of hybrid functionals; therefore the use of optB88 appears to be a sensible alternative to calculations implying the evaluation of the Fock operator, in cases when simulations of large systems are required.

1. INTRODUCTION

In the past two decades, much progress has been reported in simulations of liquid water from first principles.¹ However, discrepancies with experimental data remain, when using calculations based on density functional theory (DFT), e.g., in describing the structural and diffusive properties of the system. One issue that is receiving much attention lately is the performance of van der Waals density functionals^{2–4} (vdW-DFs) in predicting the properties of water; these are functionals designed to account for dispersion forces, at least approximately.

van der Waals forces denote the forces between atoms or molecules in dilute gases or liquids, i.e., in the presence of negligible electronic charge overlap between the constituents. According to several authors, for example ref 5, for point-like polar molecules, one may identify three different contributions to vdW forces: (i) thermal orientation, originating from dipole–dipole interactions, first described by Keesom,^{6,7} (ii) induction, stemming from dipole-induced dipole interactions, introduced by Debye,^{8,9} and (iii) dispersion, first discussed by London,¹⁰ arising from the dynamic interactions between fluctuating dipoles. The first two contributions are present only in systems whose building blocks have a permanent dipole (and thus in water), while the third one occurs between any type of atoms or molecules, irrespective of their polarity. In all three cases, the interaction energies decay as the sixth power of the distance.

In liquid water, there is a non-negligible electronic charge overlap between first neighbor molecules, whose main interaction is through hydrogen bonding, not vdW forces. Many definitions have been given of hydrogen bonding,¹¹ including a recent one by IUPAC;¹² for the purpose of the present discussion, we loosely define a hydrogen bond (HB) as the attractive interaction

between a proton donor covalently bonded to a species X (X–H) and a proton acceptor Y.¹¹ Both charge transfer (CT) effects and electrostatic interactions contribute to hydrogen bonding. A recent analysis of the relative contribution of these two components in the water dimer has been given by Khaliullin et al.,¹³ who showed that the amount of intermolecular CT is on the order of a few millielectrons, i.e., much smaller than inferred from conventional population analysis in the past. In the liquid, most of the vdW interaction energy comes from second neighbor molecules, but a small contribution is expected also from non-hydrogen bonded configurations between first neighbors, originating from geometries where both charge overlap and CT are much smaller than in HBs. In these configurations, oxygen–oxygen distances roughly correspond to the first minimum of the oxygen–oxygen pair correlation function.

To understand the performance of semilocal and hybrid density functionals (e.g., those derived within generalized gradient approximations (GGA) such as PBE,^{14,15} BLYP,^{16,17} and PBE0¹⁸) in describing the induction and orientation contributions to vdW forces, it is useful to examine how well these functionals can account for the dipole moment (μ) of water molecules in ice and water. In the gas phase, the dipole moment of the molecule has been measured to great accuracy: it is 1.855 D,¹⁹ and most DFTs can reproduce this value within 3% (see section 3.1). The same quantity in condensed phases is not well-defined and thus cannot be measured, since it is not possible to partition in a unique way the electronic charge density between individual molecules. However, the quantity μ^2G can be derived from the measured

Received: May 13, 2011

Published: August 22, 2011

dielectric constant (ϵ_0); within linear response, $\mu^2 G = 3k_B T(\epsilon_0 - 1)/4\pi\rho$, where ρ and T are the number density and temperature of the system, respectively, and k_B is the Boltzmann constant. G is a correlation factor that accounts for angular correlations among dipoles: $G = 1 + \sum_i N_i \langle \cos \theta_i \rangle$, where N_i is the number of molecules in the i th coordination shell and $\langle \cos \theta_i \rangle$ is the average cosine of the angle between the dipole of a given molecule and a dipole in the i th coordination shell ($G = 1$ for uncorrelated dipoles). For perfectly ordered ice, $G = 3$, and thus at ambient conditions, for proton disordered ice, one expects $G \leq 3$, implying $\mu \geq 3.03$ D for $\epsilon_0 \approx 107$.²⁰ Experimental estimates for liquid water, based on X-ray form factors, yield $\mu = 2.9 \pm 0.6$.²¹ One may also estimate an effective dipole moment of molecules in ice and water from calculations of the induced dipole on a given molecule due to its environment, based on electrostatics. Such calculations were pioneered by Coulson and Eisenberg,²² who used a self-consistent induction model to obtain the dipole moment of a water molecule in ice. Using values of the quadrupole moment of the isolated molecule available at the time, and a series of approximations including the neglect of intermolecular polarizability, Coulson and Eisenberg obtained a value of 2.6 D (often misquoted as the dipole moment of the molecule in liquid water or even as an experimental result). Twenty years later, Batista et al.²³ used a model similar to that of ref 22, but they employed refined values of the multipole moments of the isolated molecule and computed a value of 3.09 D, which is most probably a lower bound to the dipole moment in ice.

Within a first principle approach based on ab initio molecular dynamics, where both ice and water are considered as condensed systems, one may define an effective dipole moment from maximally localized Wannier functions²⁴ (MLWFs). These are derived from linear combinations of Bloch states, obtained by self-consistent solutions of the Kohn–Sham equations. Four doubly occupied MLWFs may be associated with each molecule (there are eight valence electrons per molecule) and $\vec{\mu}$ is computed as²⁵ $\vec{\mu} = q\vec{r}_{H_1} + q\vec{r}_{H_2} + 6q\vec{r}_O - 2q\sum_{c=1}^4 \vec{r}_{w,c}$ where $\vec{r}_{w,c} = 1$ to 4 are the four maximally localized Wannier centers (MLWCs). On average, the overlap between MLWFs of adjacent molecules is less than 1% of the norm.²⁵ When using experimental equilibrium density, the PBE approximation yields a dipole moment of water equal to 3.24 ± 0.31 D and 3.09 ± 0.34 D at ~ 325 K and ~ 439 K,²⁶ respectively, and a dipole moment of ice Ih equal to 3.32 D at 273 K.²⁷ The PBE0 approximation yields slightly lower values (3.09 ± 0.28 D at ~ 330 K and 2.94 ± 0.30 D at ~ 438 K²⁶). These all appear to be physically sound figures, indicating that induction and orientation forces in water and ice should be described in a reasonable manner by these DFTs.

The description of dispersion forces is much more delicate and implies the ability to account for complex electronic correlation effects. The contribution of dispersion forces to the binding of water is expected to be small compared to that of hydrogen bonding. For example, using the formula $C_6 R^{-6}$ with C_6 taken from the recent estimates of Grimme et al.²⁸ and R equal to the average second neighbor distance (as from the second peak of the experimental $g_{OO}(R)$), and assuming that the largest contribution to dispersion comes from the second neighbors (about 10 on average in liquid water), one obtains ~ 0.01 eV; this interaction energy is ~ 20 times smaller than the HB energy in the water dimer (0.24 eV²⁹). Although weak, dispersion forces may influence the fine details of the potential energy surface of liquid water and possibly its equilibrium density.

Attempts to include dispersion forces in ab initio simulations of liquid water fall into two categories: one strategy consists of adding $C_6 R^{-6}$ dispersion-like terms or dispersion-corrected atom-centered potentials (DCACP) to the potential energy surface obtained with semilocal DFTs:³⁰ $E_{\text{DFT-D}} = E_{\text{DFT}} + E_{\text{disp}}$.^{31,32} Another strategy is based on the use of so-called vdW-DFs² designed to include, in an approximate manner, correlation effects responsible for London dispersion forces. We note that vdW-DFs not only include a partial description of dispersion forces but also provide different results for induction and orientation terms, with respect to PBE and PBE0, as well as for hydrogen bonding and intramolecular covalent bonding. Therefore, a comparison of results obtained with GGA functionals and with hybrid functionals and those obtained with vdW-DFs does not provide a direct measure of the importance of dispersion forces in water. (Unfortunately, much confusion is present in this issue in the literature.) Schemes adding $C_6 R^{-6}$ and higher terms onto the energy obtained with GGA functionals could in principle provide a clearer separation between dispersion and HB contributions to the binding of liquid water; however, such a distinction ultimately relies on the choice of a cutoff distance at which a vdW term is added to the Hamiltonian, and this choice is by no means straightforward, especially in molecular dynamics simulations.

Recently, it has been shown^{33,34} that the use of the empirical correction proposed by Grimme together with the BLYP functional (BLYP-D) gives a better agreement with experimental results for the computed equilibrium density of the liquid (0.992 g/cm³) and the computed melting temperature (~ 360 K instead of ~ 400 K) at the experimental equilibrium density. However, the second peak of the oxygen–oxygen pair correlation function appears to be washed out,^{34,35} and the estimated oxygen coordination number is very large (7.1).³⁵ The average number of HBs is instead reasonable, ~ 3.45 .³⁵ The use of a DCACP in conjunction with the BLYP functional gives a better agreement with experimental results in terms of correlation functions and 4.6, 3.61, and 2.91 ± 0.28 D for the oxygen coordination number, average number of HBs, and dipole moment³⁵ at ~ 325 K, respectively. The use of self-consistent polarization density functional theory (SCP-DFT) proposed by Murdachaew et al.³⁶ gives instead an oxygen–oxygen pair correlation function slightly more overstructured than that with BLYP and a slightly larger oxygen coordination number (4.4 vs 4.1). However, harmonic vibrational frequencies and interaction energies of water clusters are in better agreement with experimental results than those obtained with BLYP.

The performance of some vdW-DFs for the structure and density of liquid water was recently addressed by Wang et al.,³⁷ using localized basis sets and the SIESTA code. The use of the vdW-DF originally proposed by Dion et al.² with the choice of the PBE exchange functionals appears to give a softening of the liquid structure with respect to PBE and to experimental results, and a significant improvement on the equilibrium density and diffusivity of the liquid at room temperature (1.13 g/cm³ and 0.208 Å²/ps to be compared with experimental values 1.11 g/cm³ and 0.18 Å²/ps of heavy water). However, if the revPBE exchange is adopted in the definition of vdW-DF, in ref 37, a collapse of the second coordination shell of the liquid is observed.

In this paper, we present an assessment of the performance of several vdW-DFs recently proposed in the literature to describe liquid water at the experimental equilibrium density using pseudo-potentials and plane wave basis sets. We show that the vdW-DF recently proposed by Klimeš et al.⁴ gives results in better

agreement with experimental results for the structural properties of water than the vdW-DFs originally proposed by Lee et al.³⁸ and Wang et al.³⁷ and of basically the same quality for the vibrational properties. Hereafter, we refer to the functional introduced in ref 4 as optB88; all functionals are defined in detail in the following section. The quality of the agreement with experimental results for optB88 is superior to the one obtained with PBE; however, it is slightly inferior to that obtained with PBE0 for the vibrational properties. This indicates that the optB88 functional may provide a promising framework to study water solvation properties of ions and water in contact with surfaces when large supercells are required.

The rest of the paper is organized as follows: in section 2, we describe our theoretical and computational frameworks, and in section 3, we present our results for the water molecule, dimer, and liquid water. Our summary and conclusions are given in section 4.

2. METHODS

We performed calculations of the structural and vibrational properties of the water molecule and dimer and first principles molecular dynamics (MD) simulations of 32 heavy water molecules at a fixed density of 1.108 g/cm³ using the Qbox code.³⁹ We compared the results obtained with semilocal (PBE) and hybrid (PBE0) functionals with those of four different vdW-DFs, which we denote by the following acronyms: DRSL, DRSLPBE,² LMKLL,³⁸ and optB88;⁴ these are defined in detail below. In vdW-DFs, the exchange correlation energy is defined as

$$E_{xc} = E_x^{GGA} + E_c^{LDA} + E_c^{nl} \quad (1)$$

where E_x^{GGA} is the exchange energy as defined for GGA functionals, E_c^{LDA} is the local correlation energy obtained within the local density approximation (LDA), and E_c^{nl} is the nonlocal correlation energy, expressed as

$$E_c^{nl} = \frac{1}{2} \int d^3 \vec{r} \int d^3 \vec{r}' n(\vec{r}) \phi(\vec{r}, \vec{r}') n(\vec{r}') \quad (2)$$

The kernel ϕ is a universal function² which depends on \vec{r} and \vec{r}' through the variables $d = |\vec{r} - \vec{r}'|q_0(\vec{r})$ and $d' = |\vec{r} - \vec{r}'|q_0(\vec{r}')$, and it must be evaluated numerically. $q_0(\vec{r})$ is the central quantity of vdW-DFs; it is a function of the density and the density gradient, given by

$$q_0(\vec{r}) = k_F(\vec{r}) \frac{\epsilon_{xc}^0(\vec{r})}{\epsilon_{xc}^{LDA}(\vec{r})} \quad (3)$$

In eq 3, $k_F(\vec{r}) = [3\pi^2 n(\vec{r})]^{1/3}$ is the local Fermi wave vector and

$$\epsilon_{xc}^0(\vec{r}) \approx \epsilon_{xc}^{LDA}(\vec{r}) - \epsilon_{xc}^{LDA}(\vec{r}) \left[\frac{Z_{ab}}{9} \left(\frac{\nabla n(\vec{r})}{2k_F(\vec{r})n(\vec{r})} \right)^2 \right] \quad (4)$$

The four vdW-DFs used here differ by the choice of E_x^{GGA} in eq 1 and by the screening factor Z_{ab} in eq 4. Three of the functionals adopted here contain the same E_c^{nl} as originally proposed in ref 2, except for LMKLL, where the screening factor Z_{ab} is changed from -0.8491 to -1.887 . The E_x^{GGA} is different in each functional. In the original vdW-DF (DRSL) proposed in refs 2 and 40, E_x^{GGA} is computed with the revPBE⁴¹ exchange functional. The recently proposed functionals DRSLPBE and LMKLL replace the revPBE exchange functional with the PBE and

Table 1. Selected Properties of the Water Molecule, As Computed with Different Density Functionals^a

	ν_1	ν_2	ν_3	r_{OH}	$\angle HOH$	μ
PBE	3692	1592	3801	0.972	104.3	1.811
PBE0	3834	1638	3946	0.961	104.8	1.861
DRSL	3663	1618	3769	0.971	104.5	1.801
DRSLPBE	3661	1608	3767	0.972	104.5	1.807
LMKLL	3636	1623	3741	0.971	104.8	1.800
optB88	3663	1604	3770	0.973	104.4	1.813
PBEPBE ^b	3702	1601	3804	0.971	104.1	1.813
PBE1PBE ^b	3862	1643	3965	0.959	104.8	1.862
Expt.	3832 ^c	1648 ^c	3943 ^c	0.957 ⁵¹	104.5 ⁵¹	1.855 ¹⁹

^a ν denotes vibrational frequencies in cm⁻¹. ν_1 is the symmetric stretching, ν_2 is the bending, and ν_3 is the asymmetric stretching frequency. r_{OH} , $\angle HOH$, and μ are the bond length in Å, bond angle in degrees, and molecular dipole moment in Debye, respectively. Calculations were carried out in a cubic cell with $L = 30$ Bohr and a kinetic energy cutoff of 100 Ry. The converged value of the dipole moment was obtained by linear extrapolation of results obtained with cells of $L = 30, 40, 50, 60,$ and 70 Bohr, as a function of $1/L$. ^b All electron calculations. ^c Experimental harmonic frequencies.⁵¹

PW86^{42,43} exchange, respectively. In the optB88 functional, an alternative optimized exchange functional B88¹⁶ is proposed; the exchange enhancement factor of the B88 functional originally proposed by Becke¹⁶ is

$$F_x^{B88}(s) = 1 + \mu s^2 / (1 + \beta s \operatorname{arcsinh}(cs)) \quad (5)$$

where $c = 2^{4/3}(3\pi^2)^{1/3}$, $\mu \cong 0.2743$, and $\beta = 9\mu(6/\pi)^{1/3}/2c$. Klimeš et al.⁴ suggested to modify the ratio μ/β so as to optimize the binding energies of the molecules belonging to the S22 data set, as obtained using CCSD(T) calculations. Such an optimal ratio turns out to be $\mu/\beta = 1.2$ with $\mu = 0.22$, yielding ~ 10 meV mean absolute deviation for the interaction energies of the S22 data set. The calculation of the nonlocal correlation functional in the Qbox code was carried out by transforming the costly double integral of the kernel function ϕ into an efficient direct sum of Fourier coefficients as proposed in ref 44. Additional details of the implementation, including an efficient handling of the divergence of ϕ at the origin, will be given elsewhere.

We used plane wave basis sets and norm-conserving pseudopotentials (of the HSCV type^{45,46}) with a kinetic energy cutoff of 85 Ry. Simulations were carried out with a time step of 10 au (0.24 fs) in the NVE ensemble, within the Born–Oppenheimer approximation, and trajectories were collected for 20 ps for each run. The electronic contributions to the molecular dipole moment were computed using MLWFs, evaluated at each MD step with the algorithm proposed in ref 47. The IR absorption coefficient per unit length was obtained within linear response theory from the Fourier transform of the time correlation function of the system's dipole moment:⁴⁸

$$\alpha(\omega) = \frac{2\pi\omega^2\beta}{3cVn(\omega)} \int_{-\infty}^{\infty} dt e^{-i\omega t} \langle \sum_{ij} \vec{\mu}_i^*(0) \cdot \vec{\mu}_j^*(t) \rangle \quad (6)$$

where $n(\omega)$ is the refractive index, V is the volume, $\beta = 1/k_B T$ is the inverse temperature, and $\vec{\mu}_i$ is the molecular dipole moment. In the next section, we discuss the performance of the vdW-DFs defined above for liquid water, after a brief discussion of the results obtained for the isolated water molecule and the dimer.

Table 2. Vibrational Frequencies of the Water Dimer (H₂O)₂^a

	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	d_{OO}	r_{OH}^{D}	$\angle \text{O} \cdots \text{OH}$
PBE	3686	1591	3792	3526	1610	3769	2.895	0.982	6.2
PBE0	3838	1658	3946	3707	1637	3925	2.886	0.970	5.4
DRSLL	3662	1631	3764	3589	1616	3744	3.022	0.976	5.6
DRSLLPBE	3656	1606	3757	3546	1622	3736	2.924	0.980	6.6
LMKLL	3634	1619	3735	3543	1635	3712	2.967	0.978	5.9
optB88	3638	1601	3760	3529	1619	3737	2.909	0.982	5.8
PBEPBE ^b	3695	1600	3795	3536	1621	3773	2.899	0.981	6.3
PBE1PBE ^b	3853	1644	3952	3720	1665	3935	2.896	0.968	5.9
Expt.	3797 ^c	1653 ^c	3899 ^c	3718 ^c	1669 ^c	3881 ^c	2.976 ⁵²	-	6 ± 20 ⁵²

^a ν denotes vibrational frequencies in cm⁻¹. ν_1 is the symmetric stretching of the acceptor. ν_2 is the bending of the acceptor. ν_3 is the asymmetric stretching of the acceptor. ν_4 is the symmetric stretching of the donor. ν_5 is the bending of the donor, and ν_6 is the asymmetric stretching of the donor. d_{OO} is the oxygen–oxygen distance in Å. r_{OH}^{D} (Å) is the OH bond length of the donor. $\angle \text{O} \cdots \text{OH}$ (degree) is the angle between the oxygen–oxygen distance and the OH bond of the donor. Calculations were carried out in a cubic cell with $L = 60$ Bohr and a kinetic energy cutoff 100 Ry. ^b All electron calculations. ⁴⁹ ^c Experimental harmonic frequencies.⁵³

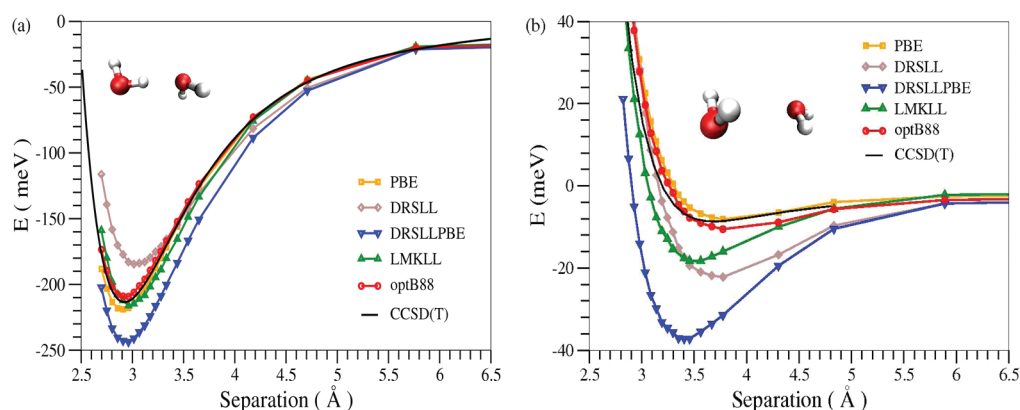


Figure 1. Interaction energy of the water dimer as a function of the separation of the water molecule center of mass, calculated with the DRSLL (gray), DRSLLPBE (blue), LMKLL (green), optB88 (red), and PBE functionals (orange). (a) Hydrogen-bonded configuration. The CCSD(T) curve⁵⁴ is shown by the black line. (b) Non-hydrogen-bonded configuration. The CCSD(T) curve, shown in black, is extrapolated to the complete basis set limit following the same procedure described in ref 55. All plane-wave calculations were carried out in a cubic cell with $L = 60$ Bohr and a kinetic energy cutoff of 100 Ry.

3. RESULTS AND DISCUSSION

3.1. Water Molecule and Water Dimer. Table 1 shows the results obtained for the structural and vibrational properties of the isolated water molecule with different functionals. PBE0 is superior to all other functionals and yields the best results for the molecule, as compared to experimental results. In particular, the average error in vibrational frequencies is 0.3%, smaller than the one obtained with PBE (3.6%) and the vdW-DFs ($\geq 3.5\%$); the PBE0 dipole moment is in excellent agreement with experimental result while all other functionals underestimate it by at least 2.3%. The bending mode of the molecule is reproduced slightly better by the vdW-DFs than by PBE, while the opposite is true for the stretching modes. The results with the PBE and PBE0 functionals obtained here for the water molecule are in excellent agreement with those reported by Xu et al.,⁴⁹ who carried out calculations with highly accurate Gaussian basis sets. Table 2 shows the intramolecular vibrational frequencies of the dimer. We find the same trend when the functional is changed as that of the isolated molecule, except for the symmetric stretching mode of the donor that is reproduced better by vdW-DFs.

Figure 1a and b show the binding curves of a hydrogen-bonded dimer (HBD) and a non-hydrogen bonded dimer (NHBD),

respectively. For the hydrogen bonded one, we used the same dimer geometry as the one optimized in quantum chemistry (QC) calculations at the CCSD(T) level;⁵⁰ the non-hydrogen-bonded configuration was extracted from a liquid water simulation, and it is representative of oxygen–oxygen distances close to the first minimum of the $g_{\text{OO}}(r)$. PBE, optB88, and LMKLL functionals all give a binding energy and a binding curve of the HBD in good agreement with that of CCSD(T), while DRSLL and DRSLLPBE yield an underestimate and overestimate of the binding, respectively. Non-hydrogen-bonded configurations appear to be much more favored by DRSLL, with respect to hydrogen-bonded ones, than by all other functionals. PBE and optB88 functional calculations best reproduce CCSD(T) results for the NHBD, although the repulsive part of the curve is slightly stiffer than that of the CCSD(T) curve.

3.2. Liquid Water. Several properties of liquid water obtained with the various functionals described in section 3.2 are summarized in Table 3. Oxygen–oxygen pair correlation functions are given in Figure 2a and b, respectively, where it is seen that the best agreement with experimental results is obtained with the optB88 functional, in simulations at ≈ 326 K, although the differences in performance with respect to DRSLLPBE appear to be minor. In particular, although the first maximum is at a distance

Table 3. Properties of Liquid Water Computed with Several van der Waals Functionals and Semilocal and Hybrid Functionals^a

	T (K)	μ (D)	N_{Hbonds}	r_{max} (Å)	$g_{\text{OO}}(r_{\text{max}})$	r_{min} (Å)	$g_{\text{OO}}(r_{\text{min}})$	$N_{\text{Coord.}}$
DRSLLPBE	295 ± 20	2.93 ± 0.25	3.56 ± 0.8	2.80	2.77	3.34	0.71	4.34
LMKLL	291 ± 20	2.83 ± 0.23	3.42 ± 0.8	2.84	2.54	3.56	0.94	5.53
LMKLL ₂₀₀	276 ± 18	2.82 ± 0.24	3.43 ± 0.8	2.87	2.65	3.52	0.91	5.30
LMKLL	237 ± 16	2.89 ± 0.23	3.62 ± 0.7	2.86	2.98	3.38	0.80	4.57
optB88	282 ± 19	3.10 ± 0.28	3.75 ± 0.6	2.76	3.37	3.34	0.50	4.24
optB88	326 ± 24	3.00 ± 0.29	3.54 ± 0.8	2.78	2.83	3.33	0.73	4.30
PBE	297 ± 20	3.27 ± 0.31	3.84 ± 0.5	2.71	3.67	3.24	0.28	3.98
PBE0	330 ± 24	3.09 ± 0.28	3.70 ± 0.7	2.68	3.01	3.31	0.58	4.37

^a All results have been obtained with a kinetic energy cutoff of 85 Ry, except for LMKLL₂₀₀, for which 200 Ry was used. T , μ and N_{Hbonds} are the average temperature, average molecular dipole moment, and average number of hydrogen bonds, respectively. r_{max} and $g_{\text{OO}}(r_{\text{max}})$ are the position and height of the first maximum in $g_{\text{OO}}(r)$. r_{min} and $g_{\text{OO}}(r_{\text{min}})$ are the position and height of the first minimum in $g_{\text{OO}}(r)$. $N_{\text{Coord.}}$ is the oxygen coordination number. Hydrogen bonds are defined by a geometrical criterion: $d_{\text{OO}} < 3.35$ Å and $\angle \text{O} \cdots \text{OH} < 30^\circ$.

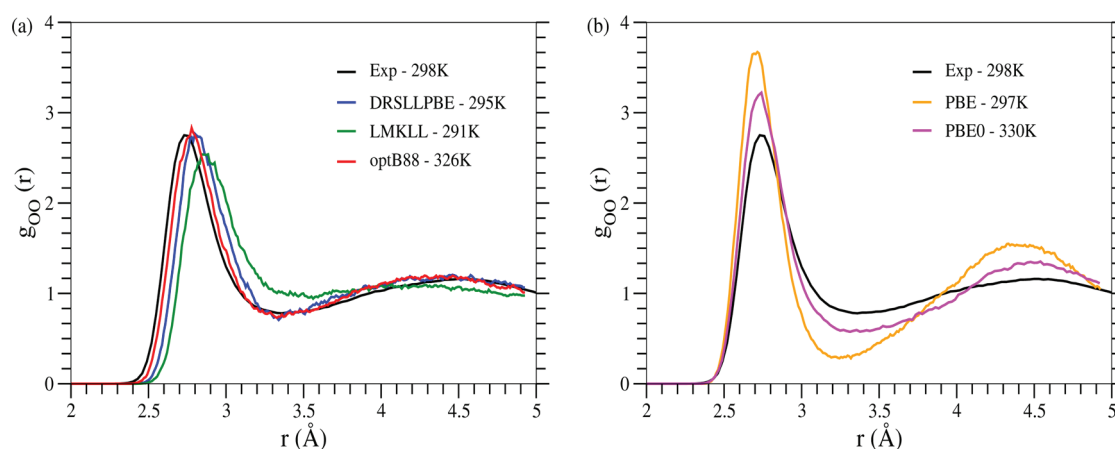


Figure 2. (a) Comparison of the oxygen–oxygen pair correlation functions calculated with vdW-DFs: DRSLLPBE at 295 ± 20 K (blue), LMKLL at 291 ± 20 K (green), and optB88 at 326 ± 24 K (red). (b) Comparison of the oxygen–oxygen pair correlation functions obtained with the PBE functional at 297 ± 20 K (orange) and the PBE0 functional at 330 ± 24 K (magenta). The experimental result⁵⁶ at room temperature is displayed by the black line in both a and b.

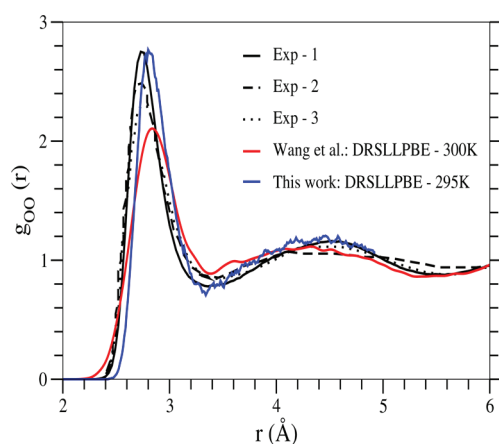


Figure 3. Comparison of the oxygen–oxygen pair correlation functions calculated with the DRSLLPBE functional in this work (blue), in ref 37 (red), and the experimental results of ref 56 (solid black), ref 57 (dashed black), and ref 58 (dotted black).

about 0.1 Å larger than in the experimental results, the position of the first minimum and the second maximum are in reasonably good agreement with those of the measured $g_{\text{OO}}(r)$ at 298 K. The

LMKLL functional gives instead the worst oxygen–oxygen pair correlation function when compared to experimental results and an unphysically large number of HBs (>5) above the experimental melting temperature (≥ 276 K); however the number of HBs is decreased to 4.57, if the temperature is lowered to ~ 237 K. A comparison of oxygen–oxygen pair correlation functions obtained with LMKLL and optB88 at different temperatures indicates that different vdW-DFs may have rather different computed melting temperatures (see Figures 1 and 2 in the Supporting Information). Our results for the diffusion coefficient are not fully converged given the MD cell sizes used here. However, they clearly indicate that the optB88 functional at 326 K has a self-diffusion constant in much better agreement with experimental results than that of simulations performed with the PBE functional at a similar temperature.

Results for the structural properties of water with the DRSLLPBE functional have also been reported in ref 37. The agreement between those results and ours is only fair, as shown in Figure 3 for the $g_{\text{OO}}(r)$. We expect these differences to originate from the use of different basis sets. We note that the accuracy of plane wave basis sets can be easily checked by varying just one parameter, as was done in the case of one of our simulations (see Table 3), where it is shown that results with 85 and 200 Ry are identical, within error bars (see Figure 1 in the Supporting

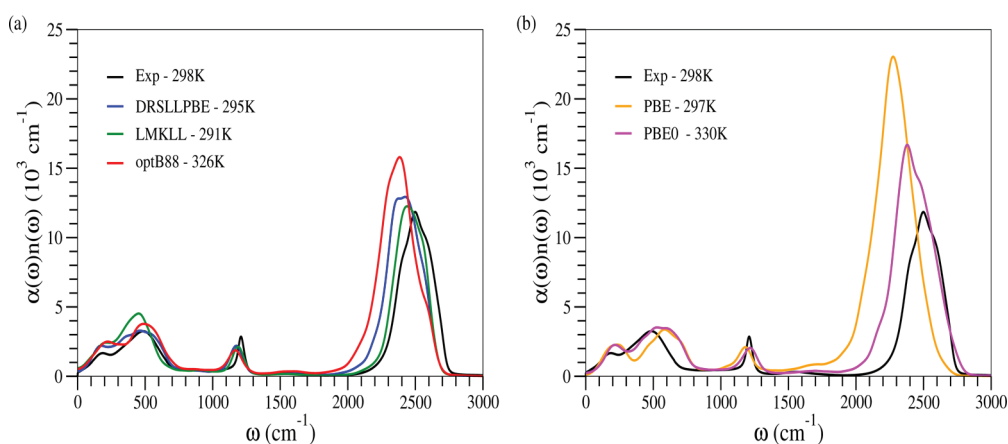


Figure 4. (a) Calculated IR spectra of liquid D_2O with vdW-DFs: DRSLPBE at 295 ± 20 K (blue), LMKLL at 291 ± 20 K (green), and optB88 at 326 ± 24 K (red). (b) Calculated IR spectra of liquid D_2O with the PBE functional at 297 ± 20 K (orange) and the PBE0 functional at 330 ± 24 K (magenta). The experimental result⁵⁹ at room temperature is displayed by the black line in both a and b.

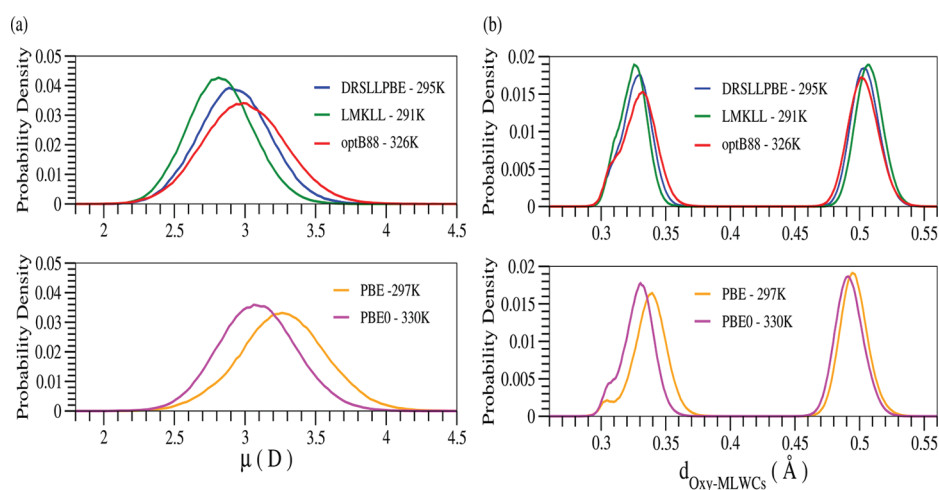


Figure 5. (a) Distributions of molecular dipole moments calculated with vdW-DFs: DRSLPBE at 295 ± 20 K (blue), LMKLL at 291 ± 20 K (green), and optB88 at 326 ± 24 K (red; upper panel) and with the PBE functional at 297 ± 20 K (orange) and the PBE0 functional at 330 ± 24 K (magenta; lower panel). (b) Distributions of distances between oxygen and maximally localized Wannier centers (MLWCs) calculated with vdW-DFs: DRSLPBE at 295 ± 20 K (blue), LMKLL at 291 ± 20 K (green), and optB88 at 326 ± 24 K (red; upper panel) and with the PBE functional at 297 ± 20 K (orange) and the PBE0 functional at 330 ± 24 K (magenta; lower panel).

Information). We have also checked our results against those of accurate Gaussian calculations for the monomer and dimer (see section 3.1) and found excellent agreement. Such a comparison for the geometry and frequencies of the isolated H_2O and $(H_2O)_2$ molecules computed with the basis sets of the SIESTA code is not available.

The IR spectra obtained with vdW-DFs appear to be superior to the spectrum computed at the PBE level and slightly superior to that computed within PBE0 for the stretching band (see Figure 4a and b). We note that the bending band is in the right position only in the PBE0 simulation. The average dipole moment in the liquid is similar when using optB88 and PBE0 (see Figure 5a), while it turns out to be sizably smaller with DRSLPBE and LMKLL. Likewise, the distributions of MLWCs are very similar when using optB88 and the hybrid functional (see Figure 5b).

We have also computed the electronic band gaps of the liquid with the different functionals. The ones computed with PBE and the vdW-DFs (obtained as an average over 22 configurations

over 20 ps) are about the same ($\approx 4.05\text{--}4.35 \pm 0.176$ eV), whereas the one obtained with PBE0 is much larger ($\approx 7.08 \pm 0.189$ eV). Although these are to be taken only as indicative values, as the cells used here are rather small and only the Γ point has been used to sample the Brillouin zone, the trend as a function of the chosen functional is significant. In particular, these results indicate that the effective polarizability of the molecule in the fluid is equally overestimated by PBE and the vdW-DFs, while much less so by PBE0. This is consistent with the values obtained for the polarizability of the isolated molecule within PBE and PBE0, 1.542 and 1.412, respectively, as compared to the experimental result of 1.427.⁴⁹

4. CONCLUSIONS

The first principles description of the properties of liquid water is an ongoing challenge, originating from the presence of several different bonding configurations which are not equally

well described by any of the known functionals. Accounting for the properties of the liquid encompasses describing with comparable accuracy intramolecular covalent/ionic bonds, intermolecular hydrogen bonds, and vdW interactions. In addition, intermolecular interactions include forces between nonbonded first neighbor molecules: these are unique bonding configurations (e.g., not present in ice at ambient conditions), with small but most likely non-negligible electronic overlap between molecular charge densities. It is clearly a formidable task to account for all of these different bonding configurations with the same level of accuracy; for example, even those functionals providing a good description of intramolecular forces and of hydrogen bonds may still fail in giving the right energy differences between hydrogen-bonded and non-hydrogen-bonded first neighbor molecules. Therefore, cancellation of errors between different configurations explored by the liquid is much more difficult to achieve in water than in a simpler system, where similar bonds are explored during time evolution. Thus, an accurate description of even the simplest properties of water, e.g., structure and diffusivity, remains elusive.

In this work, we have shown that one may find vdW-DFs of the form originally proposed by Dion et al., with appropriately chosen parameters for the local exchange, that give a description of the structural and vibrational properties of the liquid superior to that of the semilocal functional PBE. These findings indicate that the functional form for the correlation energy suggested in ref 2 is physically sound to describe water. However, our results cannot provide a direct measure of the importance of dispersion forces in liquid water. As mentioned in the Introduction, vdW-DFs do not only include an approximate description of dispersion forces (contrary to semilocal functionals that do not include any nonlocal correlations) but also give different descriptions of intramolecular and hydrogen bonds, and of the induction and orientation contributions to vdW energies, with respect to PBE.

In particular, we have shown that the optB88 functional yields the best agreement with experimental results for structure and vibrations and a liquid which displays a hydrogen-bonded network less tightly bound than so-called PBE water at the measured equilibrium density. The optB88 vdW-DF has the same form of the nonlocal correlation energy as that proposed by Dion et al.; however, the revPBE functional used in ref 2 has been substituted by the B88 exchange, reparametrized to fit the binding energies of the S22 data set, as computed with CCSD(T). Similar to the local density approximation (LDA) of DFT, where a functional form of the local correlation was chosen and fitted to the results of a highly accurate calculation (quantum Monte Carlo (QMC)) for the electron gas, the optB88 functional is derived by fitting a given functional form to highly accurate QC calculations (CCSD(T)) for molecules. It is reasonable to expect that even more accurate functionals may be derived, based on Dion et al.'s functional form, if high level calculations (either QC or QMC) become available for an ensemble of relevant systems.

The results obtained here with the optB88 functional are close to those obtained by replacing the revPBE exchange in Dion et al.'s functional with the PBE exchange. Our findings with the latter functional differ from those reported in ref 37, most likely because of the use of different basis sets. Our plane wave basis has been extensively tested for the liquid, by comparing results obtained with two different energy cutoffs, as well as for the isolated molecule and dimer, by comparing with the results of highly accurate Gaussian basis set calculations.

Although the optB88 and DRSLPBE functionals yield satisfactory results for liquid water at around room temperature, their

performance with respect to the hybrid functional PBE0 depends on the properties: oxygen–oxygen pair correlation functions are less overstructured than with PBE0, but the position of the maximum is slightly larger than in the experiment. The IR stretching bands are moderately improved, but the bending bands are inferior to that obtained with PBE0. In addition, these functionals perform worse than PBE0 for the molecule and dimer. However, the computational cost of vdW-DFs is much less than that of hybrid functionals, and therefore they are a sensible alternative to calculations implying the evaluation of the Fock operator, in cases when large systems are required.

■ ASSOCIATED CONTENT

Supporting Information. Calculated oxygen–oxygen pair correlation functions with the LMKLL and optB88 functionals for systems consisting of 32 water molecules. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gagalli@ucdavis.edu.

■ ACKNOWLEDGMENT

We would like to thank J. M. Soler, L. G. M. Pettersson, and A. Møgelhøj for useful discussions and J. Wang and M.-V. Fernández-Serra for providing some of the data of ref 37. This work was supported by the NSF through grant OCI-0749217 and grant DMR-24142900 and through TeraGrid resources provided by TACC and NICS under grant numbers TG-ASC090004 and TG-MCA06N063. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357. UC ShaRCS computational resources are also gratefully acknowledged. The CCSD(T) calculations presented in Figure 1b were carried out at the Center for Functional Nanomaterials, Brookhaven National Laboratory, which is supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under contract no. DE-AC02-98CH10886. We thank Deyu Lu for performing those calculations.

■ REFERENCES

- (1) A brief review has been given, for example, in the introduction of: Zhang, C.; Donadio, D.; Gygi, F.; Galli, G. *J. Chem. Theory Comput.* **2011**, *7*, 1443–1449.
- (2) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (3) Langreth, D. C.; et al. *J. Phys.: Condens. Matter* **2009**, *21*, 084203.
- (4) Klimeš, J.; Bowler, D. R.; Michaelides, A. *J. Phys.: Condens. Matter* **2010**, *22*, 022201.
- (5) van Oss, C. J.; Chaudhury, M. K.; Good, R. J. *Chem. Rev.* **1988**, *88*, 927–941.
- (6) Keesom, W. H. *Phys. Z.* **1921**, *22*, 129.
- (7) Keesom, W. H. *Phys. Z.* **1921**, *22*, 643.
- (8) Debye, P. *Phys. Z.* **1920**, *21*, 178.
- (9) Debye, P. *Phys. Z.* **1921**, *22*, 302.
- (10) London, F. *Z. Phys.* **1930**, *63*, 245–279.
- (11) A brief review has been given, for example, in: Buckingham, A. D.; Del Bene, J. E.; McDowell, S. A. C. *Chem. Phys. Lett.* **2008**, *463*, 1–10.
- (12) Arunan, E. *Curr. Sci.* **2007**, *92*, 17–18.

- (13) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. *Chem.—Eur. J.* **2009**, *15*, 851–855.
- (14) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (15) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (16) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (17) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (18) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (19) Shostak, S. L.; Ebenstein, W. L.; Muentner, J. S. *J. Chem. Phys.* **1991**, *94*, 5875–5882.
- (20) Johari, G. P.; Whalley, E. *J. Chem. Phys.* **1981**, *75*, 1333–1340.
- (21) Badyal, Y. S.; Saboungi, M.-L.; Price, D. L.; Shastri, S. D.; Haefner, D. R.; Soper, A. K. *J. Chem. Phys.* **2000**, *112*, 9206–9208.
- (22) Coulson, C. A.; Eisenberg, D. *Proc. R. Soc. London, Ser. A* **1966**, *291*, 445–453.
- (23) Batista, E. R.; Xantheas, S. S.; Jónsson, H. *J. Chem. Phys.* **1998**, *109*, 4546–4551.
- (24) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12865.
- (25) Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308–3311.
- (26) Zhang, C.; Donadio, D.; Gygi, F.; Galli, G. *J. Chem. Theory Comput.* **2011**, *7*, 1443–1449.
- (27) Sharma, M.; Resta, R.; Car, R. *Phys. Rev. Lett.* **2007**, *98*, 247401.
- (28) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.
- (29) Curtiss, L. A.; Frurip, D. J.; Blander, M. *J. Chem. Phys.* **1979**, *71*, 2703–2711.
- (30) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (31) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (32) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (33) Schmidt, J.; VandeVondele, J.; Kuo, I.-F. W.; Sebastiani, D.; Siepmann, J. I.; Hutter, J.; Mundy, C. J. *J. Phys. Chem. B* **2009**, *113*, 11959–11964.
- (34) Yoo, S.; Xantheas, S. S. *J. Chem. Phys.* **2011**, *134*, 121105.
- (35) Lin, I.-C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Rothlisberger, U. *J. Phys. Chem. B* **2009**, *113*, 1127–1131.
- (36) Murdachaew, G.; Mundy, C. J.; Schenter, G. K. *J. Chem. Phys.* **2010**, *132*, 164102.
- (37) Wang, J.; Román-Pérez, G.; Soler, J. M.; Artacho, E.; Fernández-Serra, M.-V. *J. Chem. Phys.* **2011**, *134*, 024516.
- (38) Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101.
- (39) Qbox code. <http://eslab.ucdavis.edu/software/qbox> (accessed May 5, 2011).
- (40) Thonhauser, T.; Cooper, V. R.; Li, S.; Puzder, A.; Hyldgaard, P.; Langreth, D. C. *Phys. Rev. B* **2007**, *76*, 125112.
- (41) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.
- (42) Perdew, J. P.; Yue, W. *Phys. Rev. B* **1986**, *33*, 8800–8802.
- (43) Murray, E. D.; Lee, K.; Langreth, D. C. *J. Chem. Theory Comput.* **2009**, *5*, 2754–2762.
- (44) Román-Pérez, G.; Soler, J. M. *Phys. Rev. Lett.* **2009**, *103*, 096102.
- (45) Pseudopotential Table. <http://fpmd.ucdavis.edu/potentials> (accessed May 5, 2011).
- (46) Vanderbilt, D. *Phys. Rev. B* **1985**, *32*, 8412–8415.
- (47) Gygi, F.; Fattiberto, J. L.; Schwegler, E. *Comput. Phys. Commun.* **2003**, *155*, 1–6.
- (48) Ramírez, R.; López-Ciudad, T.; Kumar-P, P.; Marx, D. *J. Chem. Phys.* **2004**, *121*, 3973–3983.
- (49) Xu, X.; Goddard, W. A. *J. Phys. Chem. A* **2004**, *108*, 2305–2313.
- (50) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (51) Benedict, W. S.; Gailar, N.; Plyler, E. K. *J. Chem. Phys.* **1956**, *24*, 1139–1165.
- (52) Odutola, J. A.; Dyke, T. R. *J. Chem. Phys.* **1980**, *72*, 5062–5070.
- (53) Fredin, L.; Nelander, B.; Ribbegard, G. *J. Chem. Phys.* **1977**, *66*, 4065–4072.
- (54) Molnar, L. F.; He, X.; Wang, B.; Merz, K. M. *J. Chem. Phys.* **2009**, *131*, 065102.
- (55) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. *J. Chem. Phys.* **2008**, *128*, 094313.
- (56) Soper, A. K. *Chem. Phys.* **2000**, *258*, 121–137.
- (57) Soper, A. K. *J. Phys.: Condens. Matter* **2007**, *19*, 335206.
- (58) Hura, G.; Russo, D.; Glaeser, R. M.; Head-Gordon, T.; Krack, M.; Parrinello, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1981–1991.
- (59) Max, J.-J.; Chapados, C. *J. Chem. Phys.* **2009**, *131*, 184505.

Recursive Taylor Series Expansion Method for Rigid-Body Molecular Dynamics

Alexey V. Akimov* and Anatoly B. Kolomeisky

Department of Chemistry, Rice University, Houston, Texas 77005-1892, United States

 Supporting Information

ABSTRACT: Molecular dynamics computer simulation methods are very important for understanding mechanisms of chemical, physical, and biological processes. The reliability of molecular dynamics simulations strongly depends on the integration schemes used in the simulations. In this work, we developed new rigid body integration schemes for molecular dynamics simulations. Our approach is based on a numerically exact solution to the free rigid body problem, which is used in the classical propagator splitting scheme. We use the Taylor series expansion of rotational dynamical variables in conjunction with the recursive solution for higher order derivatives of these variables. Such an approach is computationally very efficient, robust, and easy to implement, and it does not employ Jacobi elliptic functions, while still providing the numerically exact solution of the free rigid body problem. Our studies showed that the new integration methods have long-time stability and accuracy properties which are comparable to those of existing symplectic integrators. The extension to the case of a canonical ensemble is also developed, allowing one to perform simulations at constant temperatures.

1. INTRODUCTION

Molecular dynamics (MD) methods are popular theoretical tools for studying systems of various kinds, ranging from proteins,^{1–4} membranes,^{5–7} and ion channels^{8,9} to self-assembled monolayers,^{10–13} interfacial systems,^{14–17} crystalline objects,^{18,19} and biological²⁰ and artificial^{21–23} nanomachines. They rely on the classical mechanical description of the system temporal evolution as well as on the all-atomic description of the inter- and intramolecular potentials. This however implies that many interactions should be computed explicitly at every step of the MD algorithm. On the other hand, many degrees of freedom of the molecular complex under consideration might not be important for its dynamics and other properties. Therefore, in order to avoid unnecessary expensive calculations, one may combine several atoms in rigid fragments and describe the motion of such fragments in contrast to the motion of all individual atoms. Such an approach is known as rigid body molecular dynamics (RBMD). Thus, RBMD neglects some internal degrees of freedom while significantly accelerating calculations of dynamics properties of involved molecules.

In the RBMD method, the motion of rigid bodies comprising the systems is computed explicitly, including both translational and rotational (orientation) degrees of freedom. The evolution of the translational variables (position of the central of mass and the translational momentum) is usually calculated in the same way as in the all-atomic MD method using Verlet-like schemes. The most challenging part of the computation of dynamics of the rigid bodies is related to a solution for rotational variables (e.g., angular momentum and the attitude matrix).

Several RBMD methods have been developed recently.^{24–34} Most of them were constructed in such a way that they have important geometric properties of corresponding evolution operations (mapping), such as symplecticity, time reversibility, or both. The properties mentioned are usually the consequence

of the propagator construction, based on Trotter decomposition^{35,36} of the corresponding Liouville operator. It should be noted that although such a scheme always generates the time reversible mapping, it is not necessary that the mapping always be symplectic.

Using the Trotter decomposition technique, the rigid body dynamics problem is usually decomposed into two parts—the torque-free (or just free) rigid body (FRB) problem and the rest of problem, which includes the effect of forces and torques. The solution of the latter part is usually the same for most of the methods. The diversity of RBMD methods is due to various approaches to solve the FRB problem. This may be achieved by five rotations,²⁴ four rotations and the Rodrigues formula,^{25,30} rotations in quaternion space,^{25,28} as well as use of the fact that this problem is analytically solvable in terms of Jacobi elliptic functions.^{26,27,33,34} It should be noted that there are some other techniques³¹ that do not split the FRB problem from the part which includes the effect of the forces and torques.

The analytic solution of the FRB problem, utilizing the Jacobi elliptic functions, is obviously one which should be chosen due to its exact nature. However, its implementation is quite elaborate because of the necessity to consider many special cases, which may arise from one or another set of initial conditions and properties of the inertia tensors of the rigid bodies. Moreover, although there are various libraries which implement the Jacobi elliptic functions and integrals, they may introduce some code dependencies, which are not always desirable. In addition, use of such functions may require extra calculations to be preformed. On the contrary, although approximate, the methods based on five or four rotations are more easy to implement and more robust to the choice of initial conditions of properties of the rigid bodies as well as cheaper to calculate.

Received: May 17, 2011

Published: August 22, 2011

$$\begin{aligned}
 \dot{\vec{r}}_i &= \frac{\vec{p}_i}{m_i} & \dot{r}_i &= 0 \\
 \dot{\vec{p}}_i &= 0 & \dot{p}_i &= -\frac{\partial \phi}{\partial r_i} \\
 \dot{A}_i &= -\text{skew}(I_i^{-1} \vec{l}_i^b) A_i & \dot{A}_i &= 0 \\
 \dot{\vec{l}}_i^b &= \vec{l}_i^b \times I_i^{-1} \vec{l}_i^b & \dot{l}_i^b &= -\text{rot} \left(A_i^T \frac{\partial \phi}{\partial A_i} \right)
 \end{aligned}
 \tag{a} \tag{b}$$

Figure 1. Equations of motion generated by sub-Hamiltonians (a) h_1 and (b) h_2 .

Here, we report the robust and easy-to-implement method for solving the FRB problem with machine precision. It relies only on recursive relations defined by Euler equations and does not employ any special functions. This might be advantageous in many situations. It is important to note that our method might be considered simultaneously both symplectic and time-reversible. The time reversibility immediately follows from the propagator construction. Although the symplecticity does not follow from the method structure, it is a consequence of the exact nature of the solution (if properly converged).

We also extend our method as well as some of those reported recently for the case of the NVT ensemble by combining them with Nose–Poincaré thermostat.^{37–39} Finally, the case study of a water cluster is used to evaluate and compare different integrators.

2. METHOD

2.1. NVE Ensemble. One of the outstanding methods for developing the rigid body molecular dynamics integrators is a symplectic splitting technique.^{40–43} It leads to time-reversible and symplectic (usually, but not always) integrators, which are necessary for performing long-time simulations. Furthermore, one of the common splitting ways is to divide the full Hamiltonian (H) of the system of interest into a Hamiltonian of the free rigid body (h_1) and an additional Hamiltonian of interactions (h_2):

$$\begin{aligned}
 H &= h_1 + h_2 \\
 h_1 &= T(\vec{p}^N, \vec{l}^N) \\
 h_2 &= \phi(\vec{r}^N, A^N)
 \end{aligned}
 \tag{1}$$

where $T(\vec{p}^N, \vec{l}^N)$ is a kinetic energy term that depends on momenta of the rigid body centers of mass $\vec{p}^N \equiv \{\vec{p}_i\}, i \in 1-N$, and the angular moment of the rigid bodies $\vec{l}^N \equiv \{\vec{l}_i\}, i \in 1-N$, and $\phi(\vec{r}^N, A^N)$ is a potential energy term that depends on the positions of the rigid body centers of mass $\vec{r}^N \equiv \{\vec{r}_i\}, i \in 1-N$, and the orientations (attitude matrices) of the rigid bodies $A^N \equiv \{A_i\}, i \in 1-N$.

Then, the full evolution operator can be factorized: $e^{iLdt} = e^{iL_2(dt)/2} e^{iL_1 dt} e^{iL_2(dt)/2} + O(dt^3)$, where

$$\begin{aligned}
 iL &: H \rightarrow iL \\
 iL &= iL_1 + iL_2 \\
 iL &= \{\cdot, H\}, iL_1 = \{\cdot, h_1\}, iL_2 = \{\cdot, h_2\}
 \end{aligned}
 \tag{2}$$

are the Liouville operators for corresponding Hamiltonians and $\{\cdot, H\}$ denotes the Poisson bracket generated by Hamiltonian H .

In other words, it is possible to split the torqued rigid body problem into the free rigid body problem (FRB) and a perturbation term²⁴ (Figure 1).

The equations in Figure 1 can be solved separately. As one can see, the equations (Figure 1a) are the FRB problem, which can be solved in different ways. Existing methods to solve the FRB problem have been discussed in the Introduction. The symplectic splitting scheme of van Zon²⁶ is especially interesting, because it essentially provides the exact solution to this problem. However, for the correct functioning, many special cases must be treated carefully, including permutations of the axes to satisfy certain conditions. Moreover, the method relies on the set of special functions, which might be expensive to calculate, and it might lead to the additional source of rounding errors.

Another possible solution to the FRB problem is to use the Taylor expansion of both angular momentum $\{\vec{l}_i\}, i \in 1-N$, and direction vectors $\{\vec{u}_i\}, i \in 1-N$ (which specify the rigid body orientation) in the evolution time:

$$\begin{aligned}
 \vec{l}_\alpha(t + dt) &= \sum_{n=0}^N \frac{\vec{l}_\alpha^{(n)}(t)}{n!} dt^n + O(dt^{N+1}), \alpha \in x, y, z \\
 \vec{u}_\alpha(t + dt) &= \sum_{n=0}^N \frac{\vec{u}_\alpha^{(n)}(t)}{n!} dt^n + O(dt^{N+1}), \alpha \in x, y, z \\
 \vec{l}_\alpha^{(n)} &\equiv \frac{d^n \vec{l}_\alpha}{dt^n}, \vec{u}_\alpha^{(n)} \equiv \frac{d^n \vec{u}_\alpha}{dt^n}, \alpha \in x, y, z
 \end{aligned}
 \tag{3}$$

Here, we use the fact that the attitude matrix is essentially a set of three orthogonal direction vectors:

$A_{T \rightarrow e} = (\vec{u}_1 \vec{u}_2 \vec{u}_3)$. Then, the evolution of each of these vectors might be directly obtained using the rigid-body Poisson bracket:⁴⁴

$$\begin{aligned}
 \{F, G\} &= -\vec{l} \cdot (\nabla_{\vec{l}} F \times \nabla_{\vec{l}} G) \\
 &\quad - \vec{u} \cdot (\nabla_{\vec{l}} F \times \nabla_{\vec{u}} G - \nabla_{\vec{l}} G \times \nabla_{\vec{u}} F)
 \end{aligned}
 \tag{4}$$

which explicitly can be written as

$$\begin{aligned}
 \dot{\vec{u}} &= \{\vec{u}, H_1\} = \vec{u} \cdot (\nabla_{\vec{l}} H_1 \times \nabla_{\vec{u}} \vec{u}) \Leftrightarrow \\
 \dot{\vec{u}} &= \begin{pmatrix} 0 & \omega_z & -\omega_y \\ -\omega_z & 0 & \omega_x \\ \omega_y & -\omega_x & 0 \end{pmatrix} \vec{u}
 \end{aligned}
 \tag{5a}$$

$$\begin{aligned}
 \dot{\vec{l}} &= \{\vec{l}, H_1\} = -\vec{l} \cdot (\nabla_{\vec{l}} \vec{l} \times \nabla_{\vec{l}} H_1) \Leftrightarrow \begin{pmatrix} \dot{l}_x \\ \dot{l}_y \\ \dot{l}_z \end{pmatrix} \\
 &= \begin{pmatrix} \alpha l_y l_z \\ \beta l_x l_z \\ \gamma l_x l_y \end{pmatrix}
 \end{aligned}
 \tag{5b}$$

where

$$\begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = \begin{pmatrix} A l_x \\ B l_y \\ C l_z \end{pmatrix} \text{ and } \alpha = C - B, \beta = A - C, \gamma = B - A
 \tag{6}$$

Equation 5a corresponds to the equation of Figure 1a for the attitude matrix, while eq 5b is nothing else but the Euler equation for the free rigid body.

In order to use eqs 3, one must calculate the corresponding derivatives of the angular momenta as well as those for direction vectors up to the desired expansion order. This may easily be done employing the special structure of the equations of motion, which for convenience may be written as

$$\begin{aligned}\dot{l}_x &= \alpha l_y l_z \\ \dot{l}_y &= \beta l_x l_z \\ \dot{l}_z &= \gamma l_x l_y \\ \dot{u}_x &= C l_z u_y - B l_y u_z \\ \dot{u}_y &= -C l_z u_x + A l_x u_z \\ \dot{u}_z &= B l_y u_x - A l_x u_y\end{aligned}\quad (7)$$

Using the Leibniz rule, we can see that each n -th derivative of each variable (\dot{l} , $\vec{u}_1, \vec{u}_2, \vec{u}_3$) may then be expressed via derivatives of other variables up to the order of $(n-1)$:

$$\begin{aligned}(l_x)^{(n)} &= (\dot{l}_x)^{(n-1)} = \alpha \sum_{i=0}^{n-1} C_{n-1}^i (l_y)^{(i)} (l_z)^{(n-1-i)} \\ (l_y)^{(n)} &= (\dot{l}_y)^{(n-1)} = \beta \sum_{i=0}^{n-1} C_{n-1}^i (l_x)^{(i)} (l_z)^{(n-1-i)} \\ (l_z)^{(n)} &= (\dot{l}_z)^{(n-1)} = \gamma \sum_{i=0}^{n-1} C_{n-1}^i (l_x)^{(i)} (l_y)^{(n-1-i)} \\ (u_x)^{(n)} &= (\dot{u}_x)^{(n-1)} = \sum_{i=0}^{n-1} C_{n-1}^i (C l_z)^{(i)} (u_y)^{(n-1-i)} \\ &\quad - B (l_y)^{(i)} (u_z)^{(n-1-i)} \\ (u_y)^{(n)} &= (\dot{u}_y)^{(n-1)} = \sum_{i=0}^{n-1} C_{n-1}^i (-C l_z)^{(i)} (u_x)^{(n-1-i)} \\ &\quad + A (l_x)^{(i)} (u_z)^{(n-1-i)} \\ (u_z)^{(n)} &= (\dot{u}_z)^{(n-1)} = \sum_{i=0}^{n-1} C_{n-1}^i (B l_y)^{(i)} (u_x)^{(n-1-i)} \\ &\quad - A (l_x)^{(i)} (u_y)^{(n-1-i)}\end{aligned}\quad (8)$$

where $C_n^i = n!/(i!(n-i)!)$ are the binomial coefficients.

The first derivatives of all variables are calculated using the initial values of the variables themselves, as defined by eqs 7. Second derivatives may then be calculated using the first derivatives as well as the initial values of variables and so on up to the required order. The calculated derivatives then may be plugged into eqs 3 to propagate variables. The length of expansion may be chosen such that the last terms will be comparable to machine precision, which will result in a numerically exact solution of the FRB problem. We refer to the algorithm described above as Terec (TEylor REcursive, phonetically).

The evolution of the orientation of the rigid body may also be described in terms of unit quaternions. In that case, the second part of eqs 7 will read

$$\begin{aligned}\dot{q}_0 &= \frac{1}{2} (-A l_x q_1 - B l_y q_2 - C l_z q_3) \\ \dot{q}_1 &= \frac{1}{2} (A l_x q_0 - B l_y q_3 + C l_z q_2) \\ \dot{q}_2 &= \frac{1}{2} (A l_x q_3 + B l_y q_0 - C l_z q_1) \\ \dot{q}_3 &= \frac{1}{2} (-A l_x q_2 + B l_y q_1 + C l_z q_0)\end{aligned}\quad (9)$$

Similarly to direction vector eqs 8, the recursive equations for unit quaternion will read

$$\begin{aligned}(q_0)^{(n)} &= (\dot{q}_0)^{(n-1)} = \frac{1}{2} \sum_{i=0}^{n-1} C_{n-1}^i (-A l_x)^{(i)} (q_1)^{(n-1-i)} \\ &\quad - B (l_y)^{(i)} (q_2)^{(n-1-i)} - C (l_z)^{(i)} (q_3)^{(n-1-i)} \\ (q_1)^{(n)} &= (\dot{q}_1)^{(n-1)} = \frac{1}{2} \sum_{i=0}^{n-1} C_{n-1}^i (A l_x)^{(i)} (q_0)^{(n-1-i)} \\ &\quad - B (l_y)^{(i)} (q_3)^{(n-1-i)} + C (l_z)^{(i)} (q_2)^{(n-1-i)} \\ (q_2)^{(n)} &= (\dot{q}_2)^{(n-1)} = \frac{1}{2} \sum_{i=0}^{n-1} C_{n-1}^i (A l_x)^{(i)} (q_3)^{(n-1-i)} \\ &\quad + B (l_y)^{(i)} (q_0)^{(n-1-i)} - C (l_z)^{(i)} (q_1)^{(n-1-i)} \\ (q_3)^{(n)} &= (\dot{q}_3)^{(n-1)} = \frac{1}{2} \sum_{i=0}^{n-1} C_{n-1}^i (-A l_x)^{(i)} (q_2)^{(n-1-i)} \\ &\quad + B (l_y)^{(i)} (q_1)^{(n-1-i)} + C (l_z)^{(i)} (q_0)^{(n-1-i)}\end{aligned}\quad (10)$$

while the recursive relation for angular momentum will be the same as in eq 8. This version we call qTerec (quaternion Terec).

To facilitate the calculations described in MD simulations, the binomial coefficients up to a required degree may be precomputed once and for all. Also, if the length of the expansion is not long enough, the length of the direction vectors as well as that of the quaternion may change. Thus, we use the renormalization of the unit vectors (quaternion) to cure such a possible problem. It should be noted that although the renormalization of direction vectors will not solve the possible loss of the orthogonality, in the quaternion approach, this is not a problem. However, in our simulations, we have found that possible error in the orthogonality of vectors has practically no effect on dynamics and its stability and accuracy.

2.2. NVT Ensemble. All tested integration schemes (except for Omelyan³¹) can be coupled to a Nose–Poincaré thermostat in a straightforward way. The algorithm of Omelyan³¹ due to its leapfrog structure is less suitable for this purpose.

The Nose–Poincaré thermostat is introduced via an extended system Hamiltonian called the Nose–Poincaré Hamiltonian:³⁷

$$\begin{aligned}H_{\text{NP}} &= s \left[\sum_{i=1}^N \left(\frac{p_i'^2}{2m_i s^2} \right) + \sum_{i=1}^N \left(\frac{1}{2s^2} \vec{l}_i'^T I_i^{-1} \vec{l}_i' \right) \right] \\ &\quad + \phi(\vec{r}^N, A^N) + \frac{p_s^2}{2Q} + g k_B T \ln(s) - H_0\end{aligned}\quad (11)$$

where the primed letters denote the virtual variables and non-primed are real variables

$$\begin{aligned}H_0 &= H(0), H \\ &= \sum_{i=1}^N \left(\frac{p_i^2}{2m_i s^2} \right) + \sum_{i=1}^N \left(\frac{1}{2s^2} \vec{l}_i^T I_i^{-1} \vec{l}_i \right) \\ &\quad + \phi(\vec{r}^N, A^N) + \frac{p_s^2}{2Q} + g k_B T \ln(s)\end{aligned}\quad (12)$$

g is number of degrees of freedom, and k_B is Boltzmann constant.

The Hamiltonian (eq 11) may then be decomposed onto several sub-Hamiltonians:

$$\begin{aligned}
 H_{\text{NP}} &= H_0 + H_1 + H_2 + H_3 \\
 H_0 &= s \left(\sum_{i=1}^N \frac{p_i^2}{2m_i s^2} + g k_B T \ln(s) - H_0 \right) \\
 H_1 &= s \sum_{i=1}^N \left(\frac{A_i \bar{l}_{x,i}^2}{2s^2} + \frac{B_i \bar{l}_{y,i}^2}{2s^2} + \frac{C_i \bar{l}_{z,i}^2}{2s^2} \right) \\
 H_2 &= s \phi(\bar{r}^N, A^N) \\
 H_3 &= s \frac{p_s^2}{2Q}
 \end{aligned} \quad (13)$$

Every sub-Hamiltonian gives rise to its own evolution operator:

$$\begin{aligned}
 H_0(s, \bar{p}^N) \rightarrow D_0 &= \sum_{i=1}^N \left(\frac{\partial H_0}{\partial \bar{p}_i'} \frac{\partial}{\partial \bar{r}_i} - \frac{\partial H_0}{\partial \bar{r}_i} \frac{\partial}{\partial \bar{p}_i'} \right) \\
 &+ \left(\frac{\partial H_0}{\partial p_s} \frac{\partial}{\partial s} - \frac{\partial H_0}{\partial s} \frac{\partial}{\partial p_s} \right) = \sum_{i=1}^N \left(\frac{\bar{p}_i'}{m_i s} \frac{\partial}{\partial \bar{r}_i} \right) \\
 &+ \left(\sum_{i=1}^N \frac{p_i^2}{2m_i s^2} - g k_B T \ln(s) + H_0 - g k_B T \right) \frac{\partial}{\partial p_s} \\
 H_1(s, \bar{l}^N) \rightarrow D_1 &= \left(\sum_{i=1}^3 \left(\dot{\bar{u}}_i \frac{\partial}{\partial \bar{u}_i} + \dot{\bar{l}}_i \frac{\partial}{\partial \bar{l}_i} \right) \right) \\
 &+ \left(\frac{\partial H_1}{\partial p_s} \frac{\partial}{\partial s} - \frac{\partial H_1}{\partial s} \frac{\partial}{\partial p_s} \right) \\
 &= \tilde{D} + \left(\sum_{i=1}^N \left(\frac{A_i \bar{l}_{x,i}^2}{2s^2} + \frac{B_i \bar{l}_{y,i}^2}{2s^2} + \frac{C_i \bar{l}_{z,i}^2}{2s^2} \right) \right) \frac{\partial}{\partial p_s} \\
 H_2(s, \bar{r}^N, A^N) \rightarrow D_2 &= \sum_{i=1}^N \left(\frac{\partial H_2}{\partial \bar{p}_i'} \frac{\partial}{\partial \bar{r}_i} - \frac{\partial H_2}{\partial \bar{r}_i} \frac{\partial}{\partial \bar{p}_i'} \right) \\
 &+ \sum_{i=1}^N \left(\frac{\partial H_2}{\partial \bar{l}_i'} \frac{\partial}{\partial q_i} - \frac{\partial H_2}{\partial q_i} \frac{\partial}{\partial \bar{l}_i'} \right) \\
 &+ \left(\frac{\partial H_2}{\partial p_s} \frac{\partial}{\partial s} - \frac{\partial H_2}{\partial s} \frac{\partial}{\partial p_s} \right) \\
 &= \sum_{i=1}^N \left(s \bar{F}_i \frac{\partial}{\partial \bar{p}_i'} \right) \\
 &+ \sum_{i=1}^N s(\bar{r}_i) \frac{\partial}{\partial \bar{l}_i'} - \phi(\bar{r}^N, q^N) \frac{\partial}{\partial p_s} \\
 H_3(s, p_s) \rightarrow D_3 &= \left(\frac{\partial H_3}{\partial p_s} \frac{\partial}{\partial s} - \frac{\partial H_3}{\partial s} \frac{\partial}{\partial p_s} \right) \\
 &= \frac{s p_s}{Q} \frac{\partial}{\partial s} - \frac{p_s^2}{2Q} \frac{\partial}{\partial p_s}
 \end{aligned} \quad (14)$$

Finally, the full propagator (evolution operator) may be represented as

$$\begin{aligned}
 e^{D dt} &= e^{D_3 dt/2} e^{D_2 dt/2} e^{D_0 dt} e^{D_1 dt} e^{D_2 dt/2} e^{D_3 dt/2} + O(dt^3) \\
 e^{D_1 dt} &= e^{D_{11} dt/2} e^{\tilde{D} dt} e^{D_{11} dt/2} + O(dt^3)
 \end{aligned} \quad (15)$$

thus leading to the second-order factorization scheme. To build the explicit integrator, we only have to define the action of every composing operator.

The action of operators in D_0 , D_2 , and D_{11} results in the translation of corresponding variables. The nontrivial operators are thus D_3 and \tilde{D} . The action of the first one was described by Nose³⁸ and may be represented as

$$\exp[D_3 dt] \begin{pmatrix} s \\ p_s \end{pmatrix} = \begin{pmatrix} s \left(1 + \frac{p_s}{2Q} dt \right)^2 \\ p_s / \left(1 + \frac{p_s}{2Q} dt \right) \end{pmatrix} \quad (16)$$

The operator \tilde{D} describes the FRB problem in scaled angular moments. Hence, the equations of motion it generates are similar to eqs 5a and 5b:

$$\begin{aligned}
 \dot{\bar{u}} &= \{ \bar{u}, H_1 \} = \bar{u} \cdot (\nabla_{\bar{l}'} H_1 \times \nabla_{\bar{u}} \bar{u}) \Leftrightarrow \\
 \dot{\bar{u}} &= \frac{1}{s} \begin{pmatrix} 0 & \omega'_z & -\omega'_y \\ -\omega'_z & 0 & \omega'_x \\ \omega'_y & -\omega'_x & 0 \end{pmatrix} \bar{u}
 \end{aligned} \quad (17a)$$

$$\begin{aligned}
 \dot{\bar{l}}' &= \{ \bar{l}', H_1 \} = -\bar{l}' \cdot (\nabla_{\bar{l}'} \bar{l}' \times \nabla_{\bar{l}'} H_1) \Leftrightarrow \begin{pmatrix} \dot{l}'_x \\ \dot{l}'_y \\ \dot{l}'_z \end{pmatrix} \\
 &= \frac{1}{s} \begin{pmatrix} \alpha l'_y l'_z \\ \beta l'_x l'_z \\ \gamma l'_x l'_y \end{pmatrix}
 \end{aligned} \quad (17b)$$

Effectively, this means that the operator $e^{\tilde{D} dt}$ is equivalent to the solution of the FRB problem for the time dt/s . This follows from the fact that if, for some operator $D = B(\partial/\partial C)$, the evolution operator action is $e^{D dt}: C(t) \rightarrow C(t + dt)$, then for operator $D' = (1/s)B(\partial/\partial C)$, the evolution operator $e^{D' dt}$ action will be $e^{D' dt} = e^{D(dt)/s}: C(t) \rightarrow C(t + (dt)/s)$. Thus, the coupling of the rigid bodies to a Nose–Poincaré thermostat is simply the application of the integrators for the NVE case for scaled time with corresponding propagation of the thermostat variables.

For convenience, we present now the full explicit integration scheme to perform RBMD simulations in the NVT ensemble.

1.

$$e^{D_3 dt/2} : \begin{pmatrix} s \\ p_s \end{pmatrix} \rightarrow \begin{pmatrix} s \left(1 + \frac{p_s}{2Q} \frac{dt}{2} \right)^2 \\ p_s / \left(1 + \frac{p_s}{2Q} \frac{dt}{2} \right) \end{pmatrix} \quad (18)$$

2.

$$e^{D_2 dt/2} : \begin{pmatrix} \bar{p}'_i \\ \bar{l}'_i \\ p_s \end{pmatrix} \rightarrow \begin{pmatrix} \bar{p}'_i + s \bar{F}_i \frac{dt}{2} \\ \bar{l}'_i + s \bar{\tau}_i \frac{dt}{2} \\ p_s - \phi(\bar{r}^N, q^N) \frac{dt}{2} \end{pmatrix} \quad (19)$$

3.

$$e^{D_{11} dt/2} : p_s \rightarrow p_s + \frac{dt}{2} \sum_{i=1}^N \left(\frac{A_i}{2s^2} \bar{l}'_{x,i} + \frac{B_i}{2s^2} \bar{l}'_{y,i} + \frac{C_i}{2s^2} \bar{l}'_{z,i} \right) \quad (20)$$

4. FRB problem propagation with the time step dt/s , e.g., Terec (see eqs 8)

5. Exactly step 3

6.

$$e^{D_0 dt} : \begin{pmatrix} \bar{r}_i \\ p_s \end{pmatrix} \rightarrow \begin{pmatrix} \bar{r}_i + \frac{\bar{p}'_i dt}{m_i s} \\ p_s + \left(\sum_{i=1}^N \frac{\bar{p}'_i{}^2}{2m_i s^2} - g k_B T (\ln(s) + 1) + H_0 \right) dt \end{pmatrix} \quad (21)$$

7. Update atomic coordinates; calculate forces and torques

8. Exactly step 2

9. Exactly step 1

3. RESULTS AND DISCUSSION

In order to test the developed integrators, we performed MD simulations of the $(\text{H}_2\text{O})_{23}$ cluster. The TIP3P⁴⁵ interaction potential was used to describe intermolecular interactions. Each water molecule was treated as a separate rigid fragment. Thus, the inclusion of intramolecular interactions was not necessary. Each simulation runs for 10^7 steps, which for a time step of 1 fs is equivalent to a 10 ns trajectory. The initial velocity distribution corresponded to a temperature of 250 K.

The algorithms were tested in three stages. In the first stage, we compared the performance of both the Terec and qTerec algorithms for different integration time steps using different expansion sizes. The methods were characterized by two quantities: the total energy trend (b quantity in eq 22a) and the total energy standard deviation ($\text{sd}(E)$ in eq 22b).

The former quantity describes the stability of the method and therefore may be a quantitative measure of symplecticity of the method. We calculated it via a linear fit of the total energy versus trajectory time (eq 22a). It should also be noted that the symplecticity of the method cannot be simply judged on the basis of its stability. For this purpose, one should consider the phase space volume preservation. However, in most cases, the good stability of the method might reflect its symplecticness. Therefore, we consider the total energy trend as the measure of symplecticity in this sense.

The latter quantity describes how much of the total energy fluctuates around its mean value, and thus it is a measure of the method precision.

$$|E(t)| = a + bt \quad (22a)$$

$$\text{sd}(E) = \sqrt{\frac{\sum_{i=1}^N (E_i - \bar{E})^2}{N}} \quad (22b)$$

$$\bar{E} = \frac{\sum_{i=1}^N E_i}{N}$$

For each method, we considered the expansion sizes of 5, 7, 10, 12, 15, and 20 terms. Terms are explained in Figure 2. For each expansion size, the simulations with time steps of 0.5, 1, 2.5, 5, and 7.5 fs have been performed. The trajectory lengths varied from 5 to 75 ns accordingly.

We found that all methods showed good stability (Figure 2a) and accuracy properties (Figure 2b). The only exception was the Terec5 method, where the expansion size was not enough to achieve machine accuracy. It is interesting to note that the qTerec5 method showed significantly better properties than its Terec5 cousin. This is a consequence of the deorthogonalization of the direction vectors during simulations in the Terec5 method. The use of a quaternion in the qTerec5 variant precludes any problems with a possible deorthogonalization of the attitude matrix, thus leading to significantly better properties.

As shown in Figure 2 for the expansions longer than five terms, the properties of the corresponding integrators are almost independent of the expansion size. This indicates that the computations converge to machine precision. In some cases, the longer expansions are in fact slightly less stable and less precise. This may be due to accumulation of the rounding errors when dealing with very small numbers. Thus, for future use or by default, we chose 10 term series expansions for our methods, that is, Terec10 and qTerec10.

All methods are stable and accurate enough for all tested time steps except for 7.5 fs. This is clearly shown in Figure 2a,b as the abrupt change of linear relations $\ln(|E|) - \ln(dt)$ and $\ln(\text{sd}(E)) - \ln(dt)$.

In the second stage, the developed integrators (namely, Terec10 and qTerec10) have been compared to existing integration schemes which we label as DLML,²⁴ NO_SQUISH,²⁸ MN,³⁰ Omelyan,³¹ and Jacobi.²⁶ The first two algorithms are known to be both symplectic and time-reversible, while the next two are not symplectic but are time-reversible. Finally, the method based on an analytical solution to the FRB problem which uses Jacobi integrals of the first kind is by construction exact, so it should in principle be both symplectic and time-reversible as well. The comparison was based on the total energy trend (Figure 3a) as well as on the standard deviation of the total energy (Figure 3b).

As the comparison in Figure 3 illustrates, the Terec and qTerec methods show the properties of the known symplectic integrators (DLML and NO_SQUISH). The nonsymplectic schemes differ significantly from symplectic ones in both stability and precision. These observations are valid for different time steps up to 5 fs. All integrators become unstable at 7.5 fs. The properties of the Jacobi integrator are very similar to those of symplectic schemes as well as to Terec10 and qTerec10

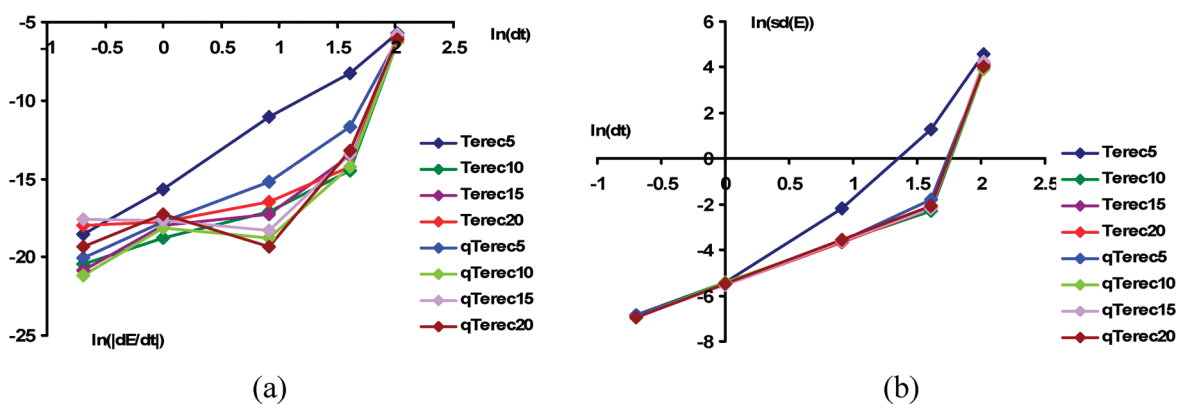


Figure 2. Comparison of the Terec and qTerec methods for different expansion sizes and different time steps. (a) energy trend, stability characteristics; (b) standard deviation, accuracy characteristics.

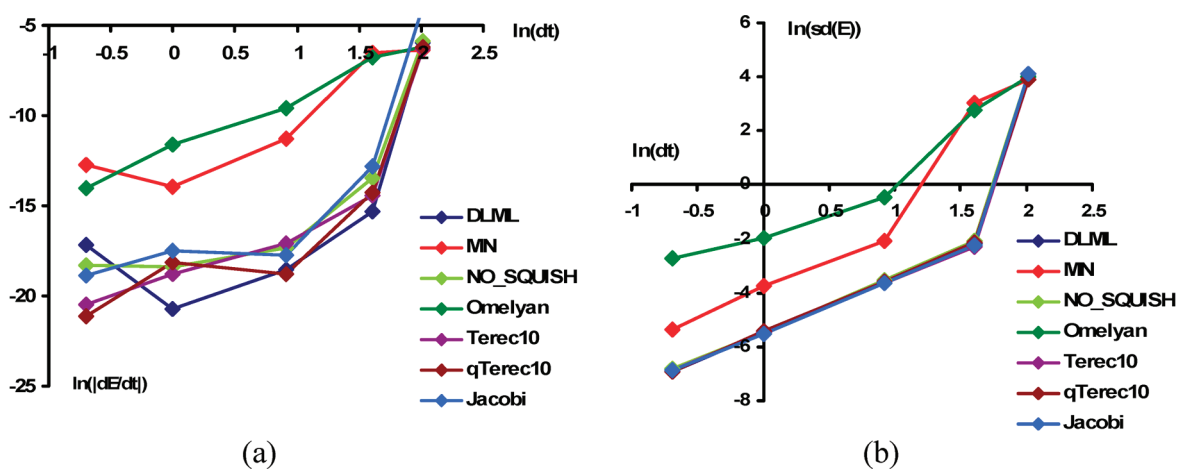


Figure 3. Comparison of Terec and qTerec methods with other existing integrators for NVE ensemble: (a) energy trend, stability characteristics; (b) standard deviation, accuracy characteristics.

algorithms. This indicates that all of them achieve the best possible precision and stability for a given system. It should also be stressed, that despite the algorithm used for integration of the FRB problem, the overall order of algorithms is limited by the smallest order in the entire decomposition scheme (that is where the forces and torques are applied). However, as is clear from Figure 3, the algorithm for solving the FRB problem may significantly impact the stability of the overall algorithms as well as its precision. Thus, it is still important to use an appropriate integrator for the FRB part.

In addition to the quality of conservation of the Hamiltonian (total energy) of the system under consideration, we also studied the analogous properties of the total linear and angular momenta of the system. This is important because the rescaling of the direction vectors and quaternions used in Terec and qTerec methods might potentially influence the conservation of these quantities. Thus, we were interested in how the rescaling affects these quantities. As we expected, the rescaling practically does not affect either the linear or angular momenta significantly, as long as the Taylor series expansion possesses a sufficiently large number of terms. This may be understood in the following way: Assume some quantity x (in our case, it may be either quaternion of the direction vector) has an exact value of x_{exact} and the Taylor series approximation of $x_{\text{Taylor}} = x_{\text{exact}} + dx$, where dx is the error.

The rescaled value of x will be $x_{\text{Taylor}}/x_{\text{exact}} = 1 + dx/x_{\text{exact}}$. As the number of terms in the Taylor series expansion increases, the error goes to zero very rapidly ($dx \rightarrow 0$). In that case, the rescaling operation will practically be the identity operation and thus will not have a significant impact on conserved properties. Corresponding data are presented in Supporting Information section S1. It shows that for a relatively small number of terms (Terec5), the rescaling affects conservation of the total linear momentum. However, starting from 10 terms (Terec10), it practically has no effect on the conserved properties and is needed only for consistency.

It should also be noted that if one wants to combine the exact solution to the FRB problem with the torques and forces part of the integrator, it is crucial to consider more than two possible ways to perform Jacobi ordering used in the Jacobi algorithm as implemented by van Zon and Schofield.²⁶ In fact, there are six distinct permutations of the axes, one of which may lead to required Jacobi ordering. However, since some of such permutations are odd, one should take particular care about the direction of time flow. The details and corresponding derivations of such modifications to the original algorithm²⁶ are presented in Supporting Information section S2.

Finally, we tested the performance of the above algorithms (except for Omelyan) in the NVT ensemble. As Figure 4 shows,

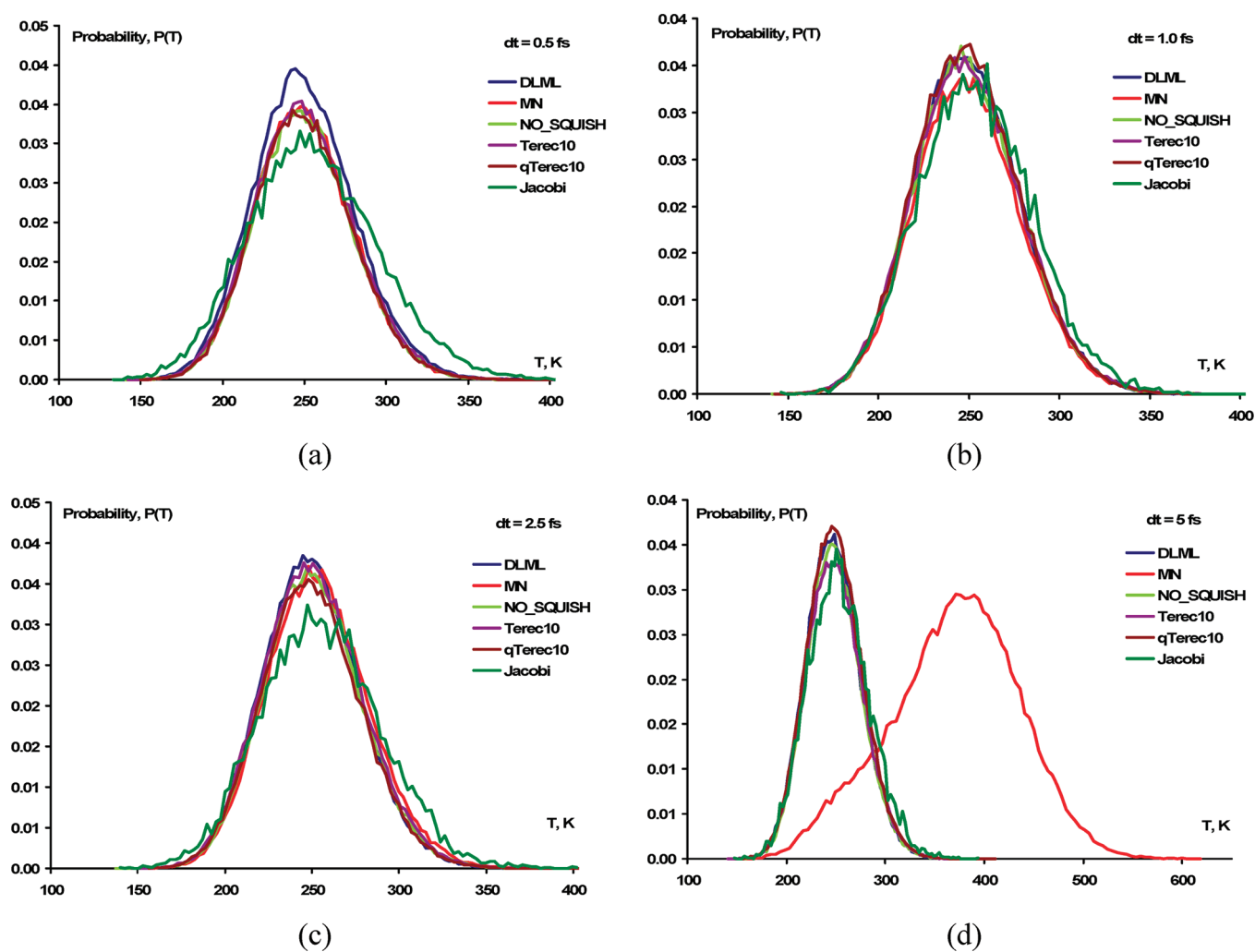


Figure 4. Temperature distribution in the NVT ensemble generated by different methods and with different integration time steps: (a) 0.5 fs, (b) 1.0 fs, (c) 2.5 fs, (d) 5.0 fs. The target temperature was set to 250 K.

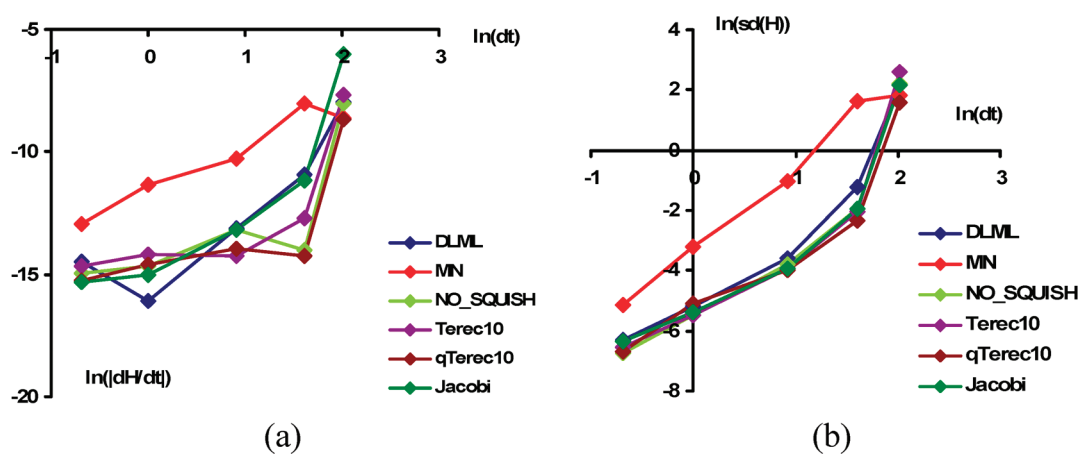


Figure 5. Comparison of Terec and qTerec methods with other existing integrators for the NVT ensemble: (a) Nose–Poincare Hamiltonian trend, stability characteristics; (b) standard deviation, accuracy characteristics.

the temperature distributions are very similar for all methods and for all time steps up to 5 fs. At $dt = 5$ fs, the MN integrator is no longer stable and does not generate the correct distribution, while other methods still perform correctly.

We also studied the stability and accuracy of the methods by examining the properties similar to those defined in eqs 22, but using Nose–Poincare Hamiltonian 11 instead of the total energy. It should be noted that the quantity (eq 11) is not only

Table 1. Run Time for a 5 ns Simulation of a Cluster of 23 Water Molecules

method	time, s
DLML	184 ± 4
MN	274 ± 5
NO_SQUISH	560 ± 4
Omelyan	92 ± 1
Terec10	373 ± 3
qTerec10	289 ± 3
Jacobi	1406 ± 70

the conserved quantity but is a true Hamiltonian. Thus, its trend gives information about the symplecticity of the NVT integrator, as the trend of total energy characterizes the symplecticity of the NVE integrators in the sense discussed earlier. The comparison of such quantities is presented in Figure 5.

Similarly to the NVE ensemble, one may observe two groups of methods—one is nonsymplectic (MN in this case), and the other group has all symplectic methods (DLML, NO_SQUISH, Terec10, and qTerec10). In contrast to the NVE ensemble, the properties of the integrator have more of an effect on the maximal integration time step for which the dynamics is still stable. As we can see, the instability occurs for the MN integrator already at $dt = 5$ fs, while such a time step is still acceptable for this method in the microcanonical ensemble. The Terec and qTerec methods once again show properties comparable to those of existing symplectic integrators.

Finally, in the third stage, the performances of all methods were compared to each other. To do this, we considered the same system, namely, a cluster of 23 water molecules, but without any interactions. This is necessary since the computation time of all interactions is much larger than the time required for performing an integration of equations of motion. That would obscure the actual speed of the integration algorithms. In other words, we considered a system of the 23 free rigid bodies, the motion of which is determined by the initial distribution of angular and linear momenta, constrained to correspond to a given temperature. In all cases, said temperature was set to 300 K. The trajectory time was set to 5×10^6 steps with an integration time step of 1 fs, which corresponds to 5 ns trajectories. It is important to note that for the purposes of a speed comparison, the trajectory length has no effect, so it could be chosen to produce any reasonable execution time.

The results of such a comparison are summarized in Table 1. The Terec method turns out to be even faster than the symplectic decomposition scheme NO_SQUISH. This is probably because the latter method uses many trigonometric functions for each integration time step. Both methods in turn are slower than the other decomposition schemes, including a quaternion version of the Terec method (qTerec). The latter is only slightly slower than the MN algorithm. However, the precision and stability comparisons made in previous stages make the MN algorithm less favorable than qTerec (and even Terec). The fastest Omelyan algorithm also suffers from stability and accuracy problems.

Finally, the only outlier is the Jacobi method. The execution time for this method becomes on average 5 times slower than that for most other algorithms. The Terec and qTerec methods also give a numerically exact solution (sometimes the precision and stability are even higher, see Figures 3 and 5), but for a

Table 2. Speed Comparisons for Different Taylor Series Expansion Sizes for Both Terec and qTerec Methods

expansion size	Terec, s	qTerec, s
5	221 ± 1	186 ± 4
10	373 ± 3	289 ± 3
15	616 ± 9	453 ± 4
20	937 ± 7	645 ± 3

fraction of the cost associated with the Jacobi method may thus be more attractive for some MD applications.

As is expected, as the expansion size increases, so do the execution times for corresponding versions of Terec or qTerec algorithms (Table 2). We can also note that for all expansion sizes, the quaternion version is usually faster than the orientation directions counterpart. This is because the first method propagates only four quaternion components, while the other propagates three components of the three direction vectors (that is, in total nine components).

4. CONCLUSIONS

In this work, we reported new numerically exact methods to solve the FRB problem that allows us to develop new integration schemes for rigid-body MD simulations. We showed how this approach might further be used for the construction of symplectic and time-reversible integrators in both microcanonical (NVE) and canonical (NVT) ensembles. Since our approach solves the FRB up to machine precision, it may be considered as an efficient and easy-to-implement alternative for existing exact solution methods, which involve Jacobi elliptic functions. Although for big integration time steps the analytic solutions might be superior, in usual MD simulations the integration time step is limited by the highest vibration frequency in the system. As a result, in most cases, it is practically impossible to use time steps larger than 5 fs, and it significantly enhances the applicability of our methods.

We showed that our integrators have characteristics not worse (but even better in some cases) than those of existing symplectic integration schemes for all time steps up to 5 fs. However, our method differs from those schemes in that it solves the FRB problem exactly (up to machine precision), and it does not need the evaluation of the Jacobi elliptic functions. Moreover, different special cases are treated in the same way as for general asymmetric rigid bodies, which facilitates the implementation of the method in computer code.

In addition, we performed a comparative study of the existing rigid body integration schemes (integrators) focusing on their stability, precision, and performance properties. We found that the time-reversible and symplectic schemes of DLML and NO_SQUISH as well as the Jacobi method based on an analytic solution of FRB show much better properties than those methods which are not symplectic (Omelyan, MN). For our new integrator, the properties depend on the expansion length. For expansion lengths as small as five terms, the integrator based on direction vectors (Terec) shows the properties comparable to tested nonsymplectic schemes, while the quaternion-based algorithms (qTerec) show much better accuracy and stability, comparable to those of symplectic schemes. For a bigger number of expansion terms, the properties of both integrators become comparable (and even better in some cases) to those of symplectic schemes. Thus, the reported methods are effectively

symplectic (which follows from the fact that the exact solution is by definition a symplectic mapping).

In terms of performance, our algorithms are much faster than the Jacobi method, providing the same and even better accuracy and stability. Moreover, our method is very robust and does not need to consider many special cases, nor deal with some internal (Jacobi) ordering. This makes them very easy to implement and efficient to run. We showed that the new methods are even faster than the existing symplectic decomposition scheme NO_SQUISH.

Although for conventional molecular dynamics the difference in performance of all algorithms is usually neglected by a significantly slower interaction calculation step, it may be more important in such methods as discrete molecular dynamics^{46–49} where the interactions are calculated relatively rarely and efficiently. In such methods, it may be necessary to solve the FRB problem for relatively long times. This may not be accomplished by conventional splitting schemes, which are approximate by construction. Using the exact method (Jacobi) described by van Zon will solve the problem, but it would take approximately 5 times more time than with our algorithms.

We also showed that the new algorithms (as well as most of the existing ones) may be combined with the Nose–Poincaré thermostat in a straightforward fashion. The corresponding mappings are similar to those used for the NVE ensemble and differ only in intrinsic scaling of the integration time step. We demonstrated that such coupling indeed generates correct distributions for all integrators considered. Moreover, our new integrators (Terec and qTerec) work in both NVT and NVE ensembles and show properties comparable to those of the existing symplectic schemes.

The advantage of our method becomes clear if one compares each of the integrators one by one. In some cases, our method is more stable and precise (vs Omelyan and MN); in others, it is faster (vs Jacobi and NO_SQUISH); in still others, it is capable of performing exact integration of the FRB problem for longer time steps (vs all splitting schemes); in some cases, it is more robust and easy to implement (vs Jacobi). We thus believe that these new integrators will be useful for long time scale simulations of various types of molecular systems.

■ ASSOCIATED CONTENT

S Supporting Information. An extension of the Jacobi method by R. van Zon is presented, giving the opportunity to perform Jacobi ordering for any combination of angular momentum components and any distinct principal components of the inertia tensor. The derivation of the corresponding transformations of the attitude matrix equation of motion is presented. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: aa7@rice.edu.

■ ACKNOWLEDGMENT

The authors acknowledge support from the Welch Foundation (Grant C-1559) and from the U.S. National Science Foundation (Grant ECCS-0708765). This work was also

supported in part by the Shared University Grid at Rice University funded by the U.S. National Science Foundation under grant EIA-0216467 and a partnership between Rice University, Sun Microsystems, and Sigma Solutions.

■ REFERENCES

- (1) Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 888.
- (2) Kuczera, K.; Jas, G. S.; Elber, R. J. *Phys. Chem. A* **2009**, *113*, 7461.
- (3) Hall, B. A.; Sansom, M. S. P. *J. Chem. Theory Comput.* **2009**, *5*, 2465.
- (4) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6679.
- (5) Jusufi, A.; DeVane, R. H.; Shinoda, W.; Klein, M. L. *Soft Matter* **2011**, *7*, 1139.
- (6) Izvekov, S.; Voth, G. A. *J. Chem. Theory Comput.* **2006**, *2*, 637.
- (7) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812.
- (8) Treptow, W.; Klein, M. L. *J. Am. Chem. Soc.* **2010**, *132*, 8145.
- (9) Zhong, Q.; Jiang, Q.; Moore, P. B.; Newns, D. M.; Klein, M. L. *Biophys. J.* **1998**, *74*, 3.
- (10) Alkis, S.; Jiang, P.; Wang, L.-L.; Roitberg, A. E.; Cheng, H.-P.; Krause, J. L. *J. Phys. Chem. C* **2007**, *111*, 14743.
- (11) Hautman, J.; Klein, M. L. *J. Chem. Phys.* **1989**, *91*, 4994.
- (12) Jung, H. H.; Won, Y. D.; Shin, S.; Kim, K. *Langmuir* **1999**, *15*, 1147.
- (13) Sellers, H.; Ulman, A.; Shnidman, Y.; Eilers, J. E. *J. Am. Chem. Soc.* **1993**, *115*, 9389.
- (14) Tierney, H. L.; Baber, A. E.; Sykes, E. C. H.; Akimov, A.; Kolomeisky, A. B. *J. Phys. Chem. C* **2009**, *113*, 10913.
- (15) Akimov, A.; Kolomeisky, A. B. *J. Phys. Chem. C* **2011**, *115*, 125.
- (16) Horinek, D.; Michl, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 14175.
- (17) Clarke, L. I.; Horinek, D.; Kottas, G. S.; Varaksa, N.; Magnera, T. F.; Hinderer, T. P.; Horansky, R. D.; Michl, J.; Price, J. C. *Nanotechnology* **2002**, *13*, 533.
- (18) Chen, J.; Trout, B. L. *J. Phys. Chem. B* **2010**, *114*, 13764.
- (19) Amirjalayer, S.; Tafipolsky, M.; Schmid, R. *Angew. Chem., Int. Ed.* **2007**, *46*, 463.
- (20) Ovchinnikov, V.; Trout, B. L.; Karplus, M. *J. Mol. Biol.* **2010**, *395*, 815.
- (21) Akimov, A. V.; Nemukhin, A. V.; Moskovsky, A. A.; Kolomeisky, A. B.; Tour, J. M. *J. Chem. Theory Comput.* **2008**, *4*, 652.
- (22) Kupchenko, I. V.; Moskovsky, A. A.; Nemukhin, A. V.; Kolomeisky, A. B. *J. Phys. Chem. C* **2011**, *115*, 108.
- (23) Konyukhov, S. S.; Kupchenko, I. V.; Moskovsky, A. A.; Nemukhin, A. V.; Akimov, A. V.; Kolomeisky, A. B. *J. Chem. Theory Comput.* **2010**, *6*, 2581.
- (24) Dullweber, A.; Leimkuhler, B.; McLachlan, R. J. *Chem. Phys.* **1997**, *107*, 5840.
- (25) Kamberaj, H.; Low, R. J.; Neal, M. P. *J. Chem. Phys.* **2005**, *122*, 224114.
- (26) van Zon, R.; Schofield, J. J. *Comput. Phys.* **2007**, *225*, 145.
- (27) van Zon, R.; Omelyan, I. P.; Schofield, J. J. *Chem. Phys.* **2008**, *128*, 136102.
- (28) Miller, T. F.; Eleftheriou, M.; Pattnaik, P.; Ndirango, A.; Newns, D.; Martyna, G. J. *J. Chem. Phys.* **2002**, *116*, 8649.
- (29) Ikeguchi, M. *J. Comput. Chem.* **2004**, *25*, 529.
- (30) Matubayasi, N.; Nakahara, M. *J. Chem. Phys.* **1999**, *110*, 3291.
- (31) Omelyan, I. P. *Phys. Rev. E* **1998**, *58*, 1169.
- (32) Okumura, H.; Itoh, S. G.; Okamoto, Y. *J. Chem. Phys.* **2007**, *126*, 084103.
- (33) Celledoni, E.; Safstrom, N. *Int. J. Model., Identification Control* **2006**, *27*, 95.
- (34) Celledoni, E.; Säfström, N. *J. Phys. A* **2006**, *39*, S463.
- (35) Trotter, H. F. *Proc. Am. Math. Soc.* **1959**, *10*, 545.
- (36) Raedt, H. D.; Raedt, B. D. *Phys. Rev. A* **1983**, *28*.

- (37) Bond, S. D.; Leimkuhler, B. J.; Laird, B. B. *J. Comput. Phys.* **1999**, *151*, 114.
- (38) Nose, S. *J. Phys. Soc. Jpn.* **2001**, *70*, 75.
- (39) Kleinerman, D. S.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. *J. Chem. Phys.* **2008**, *128*, 245103.
- (40) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. *J. Phys. A* **2006**, *39*, 5629.
- (41) Tuckerman, M. E.; Berne, B. J. *J. Chem. Phys.* **1991**, *95*, 8362.
- (42) Tuckerman, M. E.; Liu, Y.; Ciccotti, G.; Martyna, G. J. *J. Chem. Phys.* **2001**, *115*, 1678.
- (43) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990.
- (44) Leimkuhler, B.; Reich, S. *Simulating Hamiltonian Dynamics*; Cambridge University Press: Cambridge, U.K., 2004; pp 212–216.
- (45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (46) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459.
- (47) van Zon, R.; Schofield, J. J. *J. Chem. Phys.* **2008**, *128*, 154119.
- (48) de la Peña, H.; L. van Zon, R.; Schofield, J.; Opps, S. B. *J. Chem. Phys.* **2007**, *126*, 074105.
- (49) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Fold Des.* **1998**, *3*, 577.

MOIL-opt: Energy-Conserving Molecular Dynamics on a GPU/CPU System

A. Peter Ruymgaart,^{†,‡} Alfredo E. Cardenas,[†] and Ron Elber^{*,†,‡}

[†]Institute for Computational Engineering and Sciences, [‡]Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, Texas 78712, United States

S Supporting Information

ABSTRACT: We report an optimized version of the molecular dynamics program MOIL that runs on a shared memory system with OpenMP and exploits the power of a graphics processing unit (GPU). The model is of a heterogeneous computing system on a single node with several cores sharing the same memory and a GPU. This is a typical laboratory tool, which provides excellent performance at minimal cost. Besides performance, emphasis is on the accuracy and stability of the algorithm probed by energy conservation for explicit-solvent atomically detailed models. Especially for long simulations, energy conservation is critical due to the phenomenon known as “energy drift” in which energy errors accumulate linearly as a function of simulation time. To achieve long-time dynamics with acceptable accuracy, the drift must be particularly small. We identify several means of controlling long-time numerical accuracy while maintaining excellent speedup. To maintain a high level of energy conservation, SHAKE and the Ewald reciprocal summation are run in double precision. Double precision summation of real-space nonbonded interactions improves energy conservation. In our best option, the energy drift using 1 fs for a time step while constraining the distances of all bonds is undetectable in a 10 ns simulation of solvated DHFR (dihydrofolate reductase). Faster options, SHAKing only bonds with hydrogen atoms, are also very well behaved and have drifts of less than 1 kcal/mol per nanosecond of the same system. CPU/GPU implementations require changes in programming models. We consider the use of a list of neighbors and quadratic versus linear interpolation in lookup tables of different sizes. Quadratic interpolation with a smaller number of grid points is faster than linear lookup tables (with finer representation) without a loss of accuracy. Atomic neighbor lists were found most efficient. Typical speedups are about a factor of 10 compared to a single-core single-precision code.

I. INTRODUCTION

Molecular Dynamics and Its Computational Challenges.

Molecular dynamics (MD) simulations have become a vital laboratory tool for investigations of molecular processes in fundamental statistical mechanics, material science, and biophysics. Molecular dynamics trajectories provide significant insights into mechanisms, enable the rational design of new materials, and test new analytical theories by exact numerical calculations. In the present manuscript, we restrict the discussion to algorithms that follow the classical equations of motions, conserve energy, and therefore enable basic study of dynamic phenomena. Molecular trajectories are computed in small time steps using initial value solvers that propagate the solution in time steps. The existence of fast motions on the atomic scale (e.g., molecular vibrations) necessitates the use of small time steps (femtoseconds $\sim 10^{-15}$ s) that are much shorter than times of many processes of interest. For example, folding (milliseconds) and conformational transitions of proteins (microseconds) occur on time scales much longer than femtoseconds. Billions of steps (and more) must be computed in order to reach relevant times. This significant time scale gap (12 orders of magnitude from femtoseconds to milliseconds) motivates research into the extension of time scales in atomically detailed simulations.

The search for longer simulation times in atomically detailed, solvent-explicit models is theoretical, numerical, and computational. On the theory side, alternative formulations of classical

mechanics and statistical mechanics were proposed, such as the use of boundary value formulation with large time steps,¹ sampling of rare (but rapid) trajectories,² and the patching of trajectory fragments.³ These theories were aimed at reducing the number of time steps required to obtain a desired result. For specific systems, or with the acceptance of physically motivated approximations, the reduction in the overall time steps required can easily reach billions.⁴ On the numerical side, algorithms were introduced for more rapid calculations of the forces (e.g., PME⁵), and for the use of multiple time steps (e.g., RESPA⁶). The use of a small time step for fast motions and a larger time step for slower motion allows for further efficiency gains. While the impact of the theory on the computational efforts was larger than that of numerical analysis (multiple time steps as such did not increase efficiency by more than a factor of 2), there remains the interest in conducting simulations without approximations or theoretical assumptions on the system type. The third approach to long time scales is of computational techniques attached to advances in hardware. This approach has dominated the advances in straightforward simulations (no physical assumptions or constraints on the system type are assumed) in the past 10 years.

Faster computational systems have the promise of speeding up simulation times considerably. Indeed, a special purpose computing machine, the Anton,⁷ produced trajectories on millisecond time

Received: May 30, 2011

Published: August 26, 2011

scales. Cost, however, is an issue, and as a laboratory tool, Anton is out of reach of many research groups. Besides the development of a special purpose machine for MD, in the past 10 years, we have seen significant advances in parallel hardware and software architectures. While the speed of individual computing elements did not increase significantly, parallelism of different varieties is now accessible at multiple levels. Cores, or basic computing elements, are available in distributed and shared memory configurations, complicating software development but also offering new opportunities. Another new hardware development of considerable interest is the use of graphics processing units (GPU) originally designed for the game industry. The massive parallelism of the GPU attracted the attention of computational scientists, promoting the development of the high-level programming language CUDA. A number of important scientific applications were ported to the GPU platform.^{7b,8} For the cost (mid- to high-end GPU cards can be found for ~\$200), graphics processing units provide unmatched computing power. The complete computing node that was used and benchmarked in this work—an MSI-G65 board with 8 GB 1333 MHz RAM, a Phenom IIX4 965 3.4 GHz processor, a 1 TB hard drive, and a single GTX480—costs ~\$1200. This makes it possible for applications that exploit it to run at high performance speeds accessible to a broad range of investigators.

On the other hand, exploiting the promise of the GPU for molecular dynamics applications is not a trivial task and requires a departure from legacy codes, the use of new programming models, and learning to code with new constraints on memory sizes and memory hierarchy. It is therefore not surprising that general-purpose molecular dynamics software packages with modeling of explicit solvent were slow to emerge for the GPU.^{7b,8} In some cases, the performance was not satisfactory, and in others, energy conservation (essential for calculations in microcanonical ensemble) was compromised.

There are two common models for computational chemistry. The first uses high-end supercomputers, and the second employs laboratory tools such as PC and local computing clusters. Both approaches were found to be very useful throughout the years and are employed (frequently simultaneously) in many laboratories. Typically, a calculation that requires a lot of resources for a short period of time is better run at national centers, while calculations that require long-running periods are better run at laboratory resources. For example, solving hundreds of millions of coupled equations is better done using massive parallelism on hundreds of cores,⁹ while running a large number of straightforward and short trajectories to obtain statistics for thermodynamic and kinetic averages⁴ is more effectively conducted on computer clusters, common to individual laboratories.

What are the challenges in implementing codes on a combination of CPU/GPU beyond the need to learn to use a new hardware platform, its limitations, and compilers?

The first is the realization that the GPU is not optimal for everything. Though some attempts were made to put a whole MD code on a GPU, the performance is lacking for complex systems that include a diverse set of interactions (e.g., covalent and noncovalent components) or that require double precision accuracy. Memory access is important and can be a bottleneck in the calculations. If the system is simple and uniform (e.g., Lennard-Jones fluids¹⁰), then it makes sense to implement the entire MD code on the GPU. For other cases, it is not obvious.

It is therefore suggestive to use the GPU for the parts that it is best at, massive parallel execution of simple and minimal

instruction sets, and asynchronously use the cores at the CPU for complex calculations or calculations that require double precision. Practice has shown that the communication between the GPU and the CPU is not a bottleneck with systems of sufficient complexity.

The second challenge is of load balancing between the CPU and GPU. Ideally, we wish the CPU and the GPU to spend the same time on their assigned load. This is however a complex task, and so far we have not been able to achieve it in full. At present, our code still has bottlenecks in which the GPU is waiting for the CPU to complete its task. Further optimization of the load is a topic for future studies.

Third, the mixture of computer language, compilers, and libraries (OpenMP, CUDA, C, and FORTRAN) makes the code highly complex. It is difficult to maintain, and it is sensitive to hardware configurations. Relatively minor changes in hardware, operating system, or compiler versions can induce non-trivial code and compilation changes. Sustaining MD programs is simpler for those who remain in the CPU world. For the program reported in this manuscript, we offer binary executables that run on a variety of Intel/AMD multicore CPU and Nvidia CUDA-capable GPU-equipped systems with the appropriate drivers installed at <http://clsb.ices.utexas.edu/prebuilt/>. We also provide a source code but advise the user that getting CPU/GPU code to compile and produce an executable file is more difficult than typing “make” and <enter> because third party hardware and software drivers/libraries (not provided by us) are involved.

Despite the above difficulties, significant speedups were obtained on a single heterogeneous node (multicores, one GPU) for a broad range of molecular dynamics applications. The factors are ~10 compared to running on a single core with a single precision code. This makes the investments in solving the above challenges a worthwhile exercise.

In the present manuscript, we therefore describe the implementation of a molecular dynamics module of the modeling package MOIL¹¹ on such a single-node heterogeneous system. MOIL is a suite of programs that are written in FORTRAN and span a wide range of modeling tasks. Some of its options include energy minimization, reaction path calculations,¹² rate calculations by milestoneing,^{3b} molecular dynamics, and more (see <http://clsb.ices.utexas.edu/prebuilt/MOIL.pdf> for a comprehensive description). To facilitate the GPU, code was written in CUDA to compute the nonbonded list and real space nonbonded interactions on the GPU. The overall molecular dynamics driver was written in C and so were the Verlet (or RESPA) integrators. The rest of the calculations that are performed on the CPU remained in FORTRAN. Code on the CPU that influences the overall performance, the PME reciprocal summation, SHAKE¹³ of water molecules, and SHAKE of bonds that contain hydrogen atoms, was parallelized using OpenMP. We discuss below the choices made, speed of calculations, and energy conservation.

Energy conservation is especially important for calculations that determine rate and time scales from microscopic theories and modeling. While stochastic dynamics was proposed to overcome problems with energy conservation, these corrections are questionable for dynamics. For example, the use of the Langevin equation with a phenomenological friction significantly impacts the rate. At the high friction limit, the rate is inversely proportional to the friction, making it impossible to determine the rate constant from microscopic parameters only.

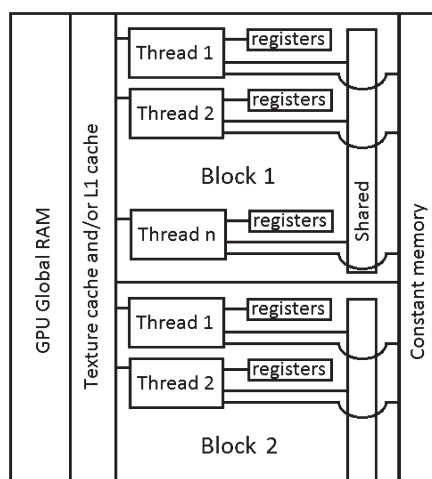


Figure 1. Conceptual memory layout of a CUDA GPU. Note the memory hierarchy in which the registers and shared memory are closest to the thread in the figure indicating fastest access. Registers are visible only to individual threads, while shared memory is accessible to a whole block of threads. Constant memory, texture cache, and L1 cache are accessible to all threads. Texture and L1 memory serve as a cache on global RAM. Data completely residing in texture cache, L1 cache, or constant memory also benefit from fast access. Note that each of these memories is small, and texture and constant memory are read-only. Direct access to global RAM is slowest, although coalescing can significantly improve its performance.

II. IMPLEMENTATION

At an abstract level, molecular dynamics algorithms typically have the following structure:

```
Do while [current number of integration steps <
desired number of integration steps]
```

1. Compute forces (bonded, real space nonbonded, reciprocal space nonbonded)
2. Increase total time by a time step
3. Compute a displacement in coordinate space
4. Adjust coordinates to satisfy constraints
5. Compute velocity displacements
6. Adjust velocities to satisfy constraints
7. Output: intermediate coordinates and velocities, energies, etc.

End while

For typical systems, the most time-consuming part is the calculation of the forces, and the most expensive component of the force calculations is the computation of the nonbonded interactions (Supporting Information D). We restrict our optimization to simulations with explicit solvation and periodic boundary conditions as the most appropriate for atomically detailed simulations of molecular kinetics (without the explicit inclusion of quantum effects). Molecular kinetics is a focus of research in our group. We split the calculation of electrostatic to real and reciprocal space using an Ewald summation scheme. Only the real part of the nonbonded calculation is performed on the GPU. The rest of the calculations are performed on the CPU using OpenMP to exploit the multicore shared memory system. Below, we provide more details on the implementation.

II.1. Memory Concerns. To appreciate the complexity of GPU programming, it is necessary to consider its memory architecture.

The GPU has a reasonably large global memory (1.5 GB in the machine which we benchmarked), which is relatively slow. Fetching data directly from the global memory to the GPU cores is costly. Therefore, programming models attempt to maximize the use of intermediate and faster memories as bridges between the global memory and the executing threads.

For a more detailed discussion on CUDA basics, we refer to previous work.^{7b,10,14} Briefly, the fastest memories available in CUDA are the thread local registers, shared memory, and constant memory. Shared memory is available only throughout a single thread block. Constant memory is available to all threads on the GPU but extremely limited in size (64 KB total) and read-only. There are few registers available per thread (~ 32), and they must be used carefully. For example, during the calculation of the nonbonded interactions between atom i and its neighbors j , thread local registers keep the coordinates of atom i , the index of this atom, and temporarily store the pair parameters for Lennard-Jones interactions A_{ij} and B_{ij} and the product of the atomic charges $q_i q_j$. The thread computes and retains (also in thread local registers) the three components of the forces operating on the i atom: f_x, f_y , and f_z . There is little space to do more on a single thread, and therefore calculations on an individual thread are conducted for one atom pair at a time (loop unrolling has little or no benefit). See Figure 1 for the conceptual memory layout of a CUDA GPU.

The next best resource after registers, constant, and shared memory are caches on global RAM. Depending on the system at hand, shared memory can be replaced by the texture memory cache or L1 cache available on the global memory of newer chips. However, as a consistent feature of GPU hardware options, old and new, the shared memory suggests itself as an indispensable bridge between the global and register memory, and we use it extensively. For example, we place the lookup table for nonbonded interactions (interpolated quadratically between the table points) on the shared memory.

II.2. Particle Meshed Ewald. Other codes port the reciprocal sum of the particle meshed Ewald (rsPME) calculations to the GPU.¹⁵ However, in our hands, the rsPME calculations in single precision were not accurate enough to avoid significant energy drift. For the purpose of the present manuscript, we define “significant energy drift” as an energy variation that exceeds 1 kcal/mol in a nanosecond simulation of DHFR, dihydrofolate reductase. Since empirically the drift is linear in time, a 1 kcal/mol drift in a nanosecond is about 1000 kcal/mol in a microsecond, enough energy to break several chemical bonds. While the excess energy is (of course) distributed in many degrees of freedom, it is likely to influence significantly the results. The kinetic energy of the system is about 15 000 kcal/mol at room temperature, and 1000 kcal/mol is $\sim 6.7\%$ of the total, or changes the temperature by ~ 20 K.

In a periodic box of water, DHFR has become a standard benchmark for molecular dynamics codes (for example, <http://ambermd.org/amber8.bench2.html>). The complexity of the rsPME calculations is considerable. About 1/3 of the total computation time in a serial code is spent in rsPME calculation, if the Verlet algorithm is used. Supporting Information D includes a profile of a serial MOIL run. Note, however, that the profile reported used RESPA,⁶ in which the reciprocal sum is computed every four steps, see below. We took the following steps to speed rsPME up: (i) It was parallelized using OpenMP. (ii) It is computed less frequently using multiple time steps. (iii) It is computed asynchronously on the CPU while the GPU is

working on real space summation. We use the reversible multiple time stepping algorithm (RESPA⁶) to integrate the equation of motion where we designate the reciprocal sum as the “slow” force. The forces are split into two, a fast varying component F^f and a slow varying force F^s . Bond constraints (SHAKE^{13a}) and matrix SHAKE for water molecules (MSHAKE^{13b}) are included in the inner loop of the algorithm.

The studies in ref 16 illustrate that the reciprocal sum has a fast force component, and a better behaving algorithm can be designed that includes a combination of real space and reciprocal space as the slowly varying force. A more elaborate and accurate force splitting has been described in which fast components of the pair interactions are removed from the reciprocal space Ewald sum. The authors reported an enhancement in the stability of the algorithm to an outer time step of 6 fs, although the level of stability depended on the choice of the parameters for a switching function used to smooth the potential energy near the cutoffs. In our hands (and also in ref 16, Figure 3), a large time step size of 4 fs for the integration of the reciprocal sum and 1 fs for the rest of the forces was stable and did not show energy drift. We therefore consider only the reciprocal force as slow. Use of this algorithmic variation in our code and its effect on stability and computation speed will be explored in the future. In the context of heterogeneous GPU/CPU computing, mixing real space and reciprocal space increases the computational complexity significantly.

The calculations of the reciprocal sum were further accelerated by parallelization of the basic FFT algorithm, and of the generation of the required charge grid using OpenMP. The parallelization was performed directly on the public domain PME code of Darden.⁵ No changes to the basic algorithm were made. We use a cubic polynomial for grid interpolation and a tolerance of error of 10^{-9} for double precision calculations. Parallelization of the non-FFT routines was done trivially by splitting up the loops between cores (e.g., in the routines fill-charge-grid and scalar-sum). The x , y , z components were parallelized with synchronization after the completion of each Cartesian direction. A recent suggestion by Schneiders et al.¹⁷ for parallelization of convolutions can improve the performance of the code. However, the present program is functional and efficient enough to remove the reciprocal sum as a major hurdle of the calculation. The bottleneck is usually still the calculation of real space nonbonded interactions.

II.3. Nonbonded List. Besides the concrete and specific implementation, perhaps the first question to ask is do we need a nonbonded list at all? In principle, there are few options: (i) do not use list, and compute all against all; (ii) use a space list based on grid partitioning; (iii) use a list based on chemical grouping; and (iv) use lists based on atoms. All of these ideas were used extensively in molecular simulations in the past, but they were reconsidered and found new ground in GPU implementations.

Option i was used for illustration purposes on systems with nonuniform particle densities. For example, protein molecules simulated with implicit solvent do not fit an exact grid and have a relatively small number of interactions. An implementation on the GPU for all-against-all interactions can provide good-looking benchmarks (results not shown). However, most MOIL applications are aimed at studies of explicit solvent systems in the NVE ensemble.

Option ii is used by a number of groups.^{7b,8,10,14} We consider a periodic system of approximately uniform density. The space is partitioned to boxes, and individual particles are placed in a box list in an operation of complexity N .¹⁸ For example, if the grid sizes along the x , y , and z directions are g_x , g_y , and g_z respectively,

a particle with coordinate x,y,z (where the origin of coordinates is such that $x, y, z \geq 0$ for all of the particles in the system) is placed in a box index $(i,j,k) = (\lfloor x/g_x \rfloor, \lfloor y/g_y \rfloor, \lfloor z/g_z \rfloor)$ where the counting starts from zero. Box neighbors are, of course, known. For each atom, there are 27 neighboring boxes, including its own box. On the basis of the box neighbor list, the interactions between atoms of nearby boxes and of the self-box are computed. The advantage of this representation is that no atomic list is produced. The atomic list tends to be long and expensive in terms of storage memory. It cannot be placed directly in the registers or in the shared memory, while box lists can reside in the shared memory. The box lists are particularly convenient for large systems, and we use them in cases that exceed 100 000 particles. A disadvantage is that box-based lists are not precise. The neighbors are not distributed spherically around each atom. The number of atom pairs is larger and requires more computational resources. Another disadvantage is that the force kernel for a box list is complex, and that hinders the use of accurate quadratic lookup tables because the limit of 32 registers per thread will be exceeded. This is because we will not be able to compactly and efficiently transfer to the thread one atom pair at a time. Yet another disadvantage is that the exclusions are not explicitly removed at the level of the lists. Therefore, when the force calculations are performed, a logical check is performed, and the excluded interactions are multiplied by zero.

Option iii of chemical grouping replaces the box-based list by a list of neighboring chemical groups. It was used in MOIL¹¹ for decades. An issue with parallelization is load-balancing, which was addressed and discussed in ref 13b. In ref 19, it was proposed to collect atoms in groups of 32 for GPU applications. This procedure provides a uniform density and fits well the GPU specifications. Nevertheless, collections of this type require periodic calculations of groups for diffusive particles (such as water molecules), and the scaling of these calculations as a function of the distances between groups is $\sim N^2$, which is unfavorable for large systems.

Option iv, which we adopt for systems smaller than 100 000 atoms, is hierarchical. We generate a box-based list, and we use that list to generate the atomic list (a list of atoms with a distance equal or smaller than a cutoff distance from the central atom i). The advantages of the atomic list are that the exclusions are taken care of during the generation of the list and that the atomic list is precise and spherical around each atom. The atomic list is generated according to two cutoffs in which the upper cutoff provides the actual list and is used as a buffer between updates of the lists. Pair calculations are performed only between pairs that are closer than the lower cutoff distance. This means that the calculations of forces are as efficient as possible. A disadvantage is the size of the list. To ensure coalesced access we follow Anderson et al.,¹⁰ the list is stored as a matrix in the global memory N_{ij} , where the index i is running over the atom number and the index j over its neighbors. It is possible to make the list more efficient by using a pointer vector, P , of the length of the number of atoms, and the actual list N . $P(i)$ points to the last neighbor of i in N .^{13b} Hence, the neighbors of i are listed between $N(i-1) + 1$ to $N(i)$. Since the density in molecular biophysics simulations is quite uniform, the number of neighbors per atom rarely changes by more than 10%, and hence both representations are of comparable complexity. For a typical real space cutoff, every atom has about 500 neighbors.

For the generation of the list for atom i , we load the index of atom i , its coordinates, and the nonbonded parameters, A_i , B_i , and

q_i , to the registers. We load from the global memory to the shared memory the coordinates and the corresponding parameters of a box neighbor to i . The individual threads can access rapidly the shared memory and produce the atomic lists of all pairs with distances lower than the distance cutoff. They also produce the corresponding products $A_{ij} \equiv A_i A_j$, $B_{ij} \equiv B_i B_j$, and $q_i q_j \equiv q_{ij}$. The preliminary calculations of the products provide more benefit than saving a few multiplications for a pair of atoms. These products are stored together in the global memory and can be accessed when needed for the force calculations in a coalesced and efficient way. The GPU has a particularly efficient memory operation (float4 array) in which four floating-point numbers are transferred between memory types in one chunk. We transfer in a float4 chunk, the neighbor index j , the Lennard-Jones parameters for the pair $A_{ij} B_{ij}$, and the charge product q_{ij} . Exclusion of bonded atoms is accomplished by chemical sorting of the atoms to begin with. Most of the bonded atoms are within 32 positions of the general atom index (with the exception of S–S bonds, prosthetic groups like heme, etc.), allowing the storage of an exclusion decision in a single bit; +1 means exclusion. The basic algorithm for exclusion is the same as that reported in ref 7b; however, the retention and preliminary calculations of the pair parameter are new and add significantly to the performance. For the rare (if any) excluded pairs that are out of the 32 range, the corresponding force is subtracted in a separate kernel.

II.4. Nonbonded Forces. We describe in detail the calculation of the forces, utilizing an atomic list. Similarly to the calculation of the list, a thread is associated with a single atom i . We store and accumulate in registers the forces that operate on this atom, and we loop over the neighboring atoms with their prepared coefficients (see section II.3). The coefficients are copied in a coalesced and efficient way from the global memory in float4 arrays. Coordinates of neighbors need to be read from global memory, and these reads are not coalesced. This however does not lead to significant performance lag, probably because the large amount of computations that follow covers up some of the memory fetch time and the coordinate fetch is at least partially cached. Similar performances are obtained if the coordinates are read from the texture cache or directly from the global memory of newer chips which feature an adjustable L1 cache on global RAM. In the latter case, we adjust the L1 cache to its maximum size of 48 KB.

Once all of the parameters are in, the square of the distance between a pair of particles is computed on each thread. With the square of the distance at hand, explicit and direct calculations of the electrostatic and van der Waals interactions can be performed. A useful trick, which has been employed for a long time now in MD codes, is the use of lookup tables to present complex functions of the distance (e.g., the error function that is used in Ewald sums). The lookup table provides the value of the function on a grid. For a more precise value, an interpolation between the grid points is required. If the table is dense (say, 10 000 grid points), the interpolation can be simple (linear), as is done in ref 7b, fetching the value from the texture memory. The texture memory is however relatively slow even when 100% cached, and it is desirable to place the table in shared or constant memory. This is possible only for small tables. We successfully used exceptionally short tables with 256 or 512 entries interpolated with quadratic lookup (see Supporting Information A and C). The accuracy of the forces is sufficient for good energy conservation, and memory access is no longer a bottleneck. The calculations of real space PME electrostatic are complex and justify the use of a table and memory fetches. We found out

however that the Lennard-Jones interactions are faster computed directly instead of looking up for values in interpolated tables. Hence, the Lennard-Jones calculations are computed directly on the thread.

Note that we do not take advantage of the symmetry in interactions between particles; i.e., we compute the interaction between the pair of atoms (ij) twice. We compute it once with atom i at the center and a second time with atom j at the center. This is clearly inefficient. However, minimizing communication is more important here than saving floating point operations. Overall, recomputing these interactions makes more sense here, as is also done by others.^{7b,8,10}

Pseudocode for the generation of the nonbonded lists and the calculations of the forces can be found in Supporting Information B and C, respectively.

II.5. Use of OpenMP. Computing environments change rapidly. A trend of the past few years is a significant increase in the number of cores on a CPU and the creation of powerful and shared memory machines. The recent announcement by Intel of a chip with 50 cores <http://www.electronista.com/articles/10/05/31/intel.knights.corner.targets.highly.parallel.pcs/> is a particularly significant step, making shared memory programming a mainstream approach. While a distributed computing approach can provide a significantly larger number of cores, the ease and efficiency of a shared memory implementation where communication speed is no longer an issue is very attractive. Shared memory machines are also more likely to be the machines of choice for laboratory instruments. We therefore decided to base our new implementation of MOIL on a model of shared memory CPU with the relatively low-cost option of a GPU. Most molecular dynamics programs (including an earlier version of MOIL^{11,13b}) were built for a distributed computing environment, and their parallelization is based on message passage interface (MPI, <http://www.mcs.anl.gov/research/projects/mpi/>). Here, we are starting fresh and base our programming model on OpenMP (<http://openmp.org/wp/>), which is a library designed for shared memory systems. We parallelized the reciprocal sum calculations of the PME (see section II.2), the Verlet algorithm, and MSHAKE algorithm (matrix SHAKE^{13b}) that constrained the geometries of water molecules. While a parallel algorithm for general SHAKE constraints was designed in the context of the MOIL program,^{13b} at present, this algorithm is not available in OpenMP. This is the topic of future work. Similarly to other groups, we parallelize SHAKE constraints for bonds that contain hydrogen atoms. We call this variant SHKL (SHAKE light). These bonds are separated into independent blocks of constraints and can be trivially parallelized on the multiple cores by OpenMP. We also comment that if SHAKE of all bonds is calculated (at present only serially), then the energy conservation with a 1 fs time step and room temperature simulation of solvated DHFR is excellent. There is no detectable drift when we use a 4096-entry quadratic-interpolation table of the real-space electrostatic interactions. However, the cost of serial calculations of SHAKE is high, and it reduces the overall speed of calculations by about 30%.

In Figure 2, we sketch the scaling of our OpenMP calculations of DHFR on a single node. An unusual observation is that the scaling of SHKL is better than the number of processors. This result is obtained since the convergence decision is made locally in each core. It is not necessary to wait for convergence for the slowest relaxing bond in the list (as is done in a serial calculation), but instead we can decide on the convergence based on the local

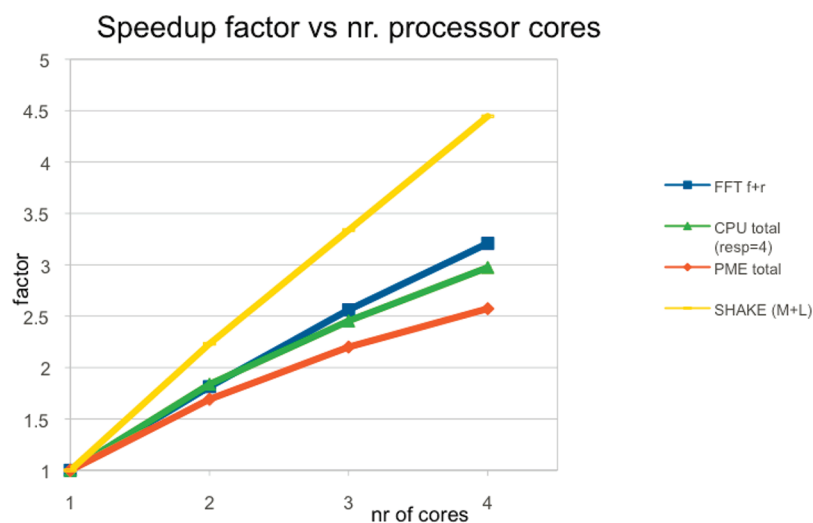


Figure 2. Efficiency of parallelization of components of the molecular dynamics algorithms computed with OpenMP. The yellow line is the scaling of the SHAKE algorithm. The red line is the complete PME reciprocal sum. The blue line is only the fast Fourier transform of the PME reciprocal sum (f is forward and r reverse), and the green line is the scaling of the overall OpenMP code. See the text for more details.

bonds assigned to a core. These bonds are decoupled from the rest. The figure also shows that further optimization of the PME code is needed and is a topic for future work.

We did not try to design an algorithm where all of the calculations are performed on the GPU (or did not try to maximize the role of the GPU in executing the calculations) due to concerns about accuracy and single precision. Calculations on the GPU are efficient if done in single precision. Recent GPU architecture allows calculations in double precision, but the penalty in performance is high. Besides the final summation of the forces, using double precision for nonbonded force calculations on the GPU is not practical at present. The reason is not only that double precision operations are slower on the GPU but also that the memory is limited. Doubling the data needed means that we will require registers that are simply not available, and the code will be unacceptably slow. Our requirement for energy conservation mandated that the reciprocal sum be computed in double precision, and therefore we prefer to execute it on the CPU. Further discussions about energy conservation can be found in section III.6.2.

III. RESULTS

All tests were run (unless specifically stated otherwise) on an MSI-G65, 8 GB 1333 MHz RAM, Phenom IIX4 965 3.4 GHZ with a single GTX480. We describe below applications of MOIL-opt to moderately sized systems. These systems are more likely to be investigated in a laboratory setting, which is our target. Since they are relatively small, they are less likely to produce impressive benchmarks. However, as we illustrate below, we are getting consistently good performance with no compromises in the numerical accuracy of the simulations. We discuss typical inputs, conditions, and timing. In the calculations below, the tolerance for relative errors of constraints was always 10^{-12} . The recommended choice for the reciprocal sum calculations on the CPU is to be conducted in double precision, and allowed relative errors were 10^{-9} . We discuss in section III.6.2 the impact of a single precision calculation. As a reference for performance, we compared these simulations (GPU code with mixed single/double

precision accuracy) to simulations conducted on one CPU core in single precision. While critical accuracy can be lost in single precision calculations, these comparisons provide a good measure and a lower bound of the speedup obtained in the heterogeneous environment of the GPU/CPU. The other alternative we have in MOIL (calculations on one core in double precision) is significantly slower. The new lookup tables built for the GPU/CPU code are more accurate than the tables of the double precision, older version of MOIL.¹¹ As a result, the energy conservation on long time scales is actually better in the GPU/CPU version than the older double precision version. Sections III.1–III.5 focus on the performance of MOIL-opt for different molecular systems.

A quick summary of our observations is given in Table 1.

III.1. DHFR. We report a simulation in a periodic box of aqueous solution of size $62.23 \times 62.23 \times 62.23 \text{ \AA}^3$ with the use of PME for long-range electrostatics. Real space cutoff distances define two layers (to create a buffer during the calculations of the forces), which are 8.5 \AA and 8.8 \AA . Forces between particles in the list that are separated by distances larger than the lower cutoff are not computed. The total number of boxes used in the generation of the nonbonded list was 7^3 .

The nonbonded list was updated every seven steps. The time step was 1 fs, and the PME mesh was of 64^3 points. The number of particles was 23 536. The force field was OPLS-AA²⁰ with the TIP3P water model.²¹ A quadratic lookup table for electrostatic interactions with 256 entries was used. Lennard-Jones interactions were computed explicitly. Water molecules were kept rigid with MSHAKE (Matrix SHAKE), and all bonds that include hydrogen atoms were constrained as well. All components of the program were run in parallel, either on the GPU or on (at most) four cores of the CPU. Direct, constant energy simulations (NVE ensemble) were conducted. The RESPA algorithm was used with a 4 fs time step for integrating the PME reciprocal forces and 1 fs for the rest of the interactions. Examining the exclusion list, only three particle pairs were beyond the 32 range for this system. The interactions of these pairs were explicitly subtracted on the GPU from the forces of these particles. The calculations on the CPU and the GPU are conducted (to the extent possible) asynchronously.

Table 1. Summary of the Computational Costs Required to Run Different Components of MOIL-opt for Different Molecular Systems^a

	Nb real	Nb recip	list gene	SHAKE	total
DHFR(23 536 atoms)	7.80 ms/step	7.19 ms/step	1.24 ms/step	2.47 ms/step	4.5 ns/day
membrane (38 802 atoms)	16.40	8.3	2.16	1.76	3.18
peptide I (20 235 atoms)	12.20	8.32	1.12	1.62	3.7
Trp zipper (5847 atoms)	2.2	1.73	0.5	0.41	45.0
peptide II (2690 atoms)	1.30	0.23	0.20	0.13	44.1

^a Only the components with highest costs are shown. The cost is expressed in milliseconds per step with the exception of the last column (nanoseconds per day). The results in this table were obtained using four cores in a CPU. For peptide I, only three cores were used. “Nb real” means the real part of the nonbonded interactions, which is computed on the GPU. “Nb recip” is the Ewald reciprocal sum which is calculated on the CPU. “list gene” is the generation of the neighbor list computed on the GPU. “SHAKE” is the enforcement of constraints, typically keeping water molecules rigid and fixing bond lengths between pairs of particles, one of which is hydrogen atom. The “total” is the number of nanoseconds produced in a day of calculation. See the text for more details on the simulations.

At the time that the nonbonded interactions are computed on the GPU, the CPU is busy computing the reciprocal sum and the covalent interactions. It should be noted that, overall, the GPU is faster. It typically waits for the CPU cores we used for the OpenMP parallelized calculations to complete their tasks. The use of more cores in the future can tilt this balance. However, more than speed, our concern with the distribution of the work between the GPU and the CPU is one of accuracy and energy conservation.

The run described above with a single thread on the CPU generates 2.27 ns a day, with two threads 3.42 ns/day, with three threads 4.07 ns/day, and with four threads 4.5 ns/day. The CPU scaling leaves room for improvement, but the use of multiple cores is still useful. Times for different components in the calculations (per step) for a single core are as follows: Real space nonbonded interactions (GPU), 7.80 ms; covalent interactions (CPU), 1.31 ms; reciprocal sum (computed every four steps on the CPU), 18.5 ms; generation of nonbonded list (GPU), 1.24 ms; SHAKE (CPU), 10.98 ms; Verlet integration (CPU), 0.50 ms. Note the significant computation time for the PME and SHAKE calculations (that can benefit from more cores).

Using four threads, the timing of the CPU changes as follows: covalent interactions (not parallel, since the overall contribution is small), 1.31 ms; reciprocal sum, 7.19 ms; SHAKE, 2.47 ms; Verlet, 0.55 ms. It is clear that adding more cores and bringing the calculation on the CPU to be on par with the timing of the GPU can speed up the calculations further.

A comparable single precision calculation on one core (no GPU) generates 421 ps per day. Hence, the speed up we see for this system is about a factor of 10, while the accuracy is higher on the mixed CPU/GPU system.

The DHFR system is a standard benchmark in the field. However, many other applications different in sizes and complexities are also investigated in numerous laboratories, and it is important to appreciate the flexibility of the code in addressing different systems. The examples below also illustrate the range of changes in the system parameters that we can do while still retaining excellent energy conservation.

III.2. Membrane System. Consider a bilayer membrane that consists of 128 DOPC phospholipids with 38 802 particles including SPC water molecules²² and sodium chloride ions. Besides the water model, in this case, the force field was OPLS-UA.²³ The system is embedded in a periodic box of $65 \times 65 \times 120 \text{ \AA}^3$. The simulation runs at room temperature with a time step of 1 fs and the NVE ensemble. The cutoff distances were 9.8 Å and 9.0 Å. Periodic boundary conditions and PME for

electrostatic calculations are used. The number of grid points was 64^3 for the mesh. There were no excluded atoms beyond the 32 range, ensuring efficient calculations of the exclusion lists and the real space nonbonded interactions. A lookup table of 256 grid values with quadratic interpolation was used to compute real-space electrostatic interactions. Water molecules were MSHAKE'ed. Simulating this system on four threads produced 3.18 ns/day. By far, the most expensive calculation is the real space nonbonded interactions (total of 16.40 ms/step). The calculation of the neighbor list requires 2.16 ms/step. Calculations conducted on the CPU with four threads include covalent forces (1.93 ms/step), reciprocal sum (8.38 ms/step), and SHAKE (1.76 ms/step).

In this particular case, it is the GPU that is slower and completes its task after the CPU. Additional savings in computer time are obtained from the asynchronous calculations, about 4 ms/step or 14.9% of the total time.

On one core, using a single precision calculation, the performance is of 209 ps a day. The GPU/CPU system speeds up these calculations by about a factor of 15, to 3.18 ns per day.

III.3. Twenty Amino Acid Solvated Helix (LKKLGKLLK-KLLKGLKLLK) (Peptide I). The motivation of studying this peptide was ref 24. We collaborated with the authors that helped us build the initial system. Both MSHAKE for water molecules and SHKL were used in parallel on the CPU. The temperature was 300 K, and the cutoff distances were 9.8 Å and 9.0 Å. The nonbonded list was updated every eight steps. The calculations were conducted in the NVE ensemble in a box of size $59 \times 59 \times 59 \text{ \AA}^3$ with a grid of 64^3 for the reciprocal space calculations. We also used the most accurate interpolation option for real space electrostatic calculation (quadratic interpolation with a table of 4096 entries). This explains why the performance on the GPU is not so impressive in this case. The time step was 1 fs. The total number of particles was 20 235. The calculations on the GPU took 12.20 ms/step for nonbonded interactions and 1.12 ms/step for the generation of the list. The reciprocal summation was the most expensive calculation conducted on the CPU (with three threads): 8.32 ms/step, followed by SHAKE (1.62 ms/step), the Verlet integration (no RESPA was used in the present example; 0.40 ms/step), and covalent interactions (0.23 ms/step). The overall speed of this simulation of a peptide solvated in a large box of water was 3.7 ns/day

The single precision run on one core in this case produced 426 ps per day. This is about a factor of 9 slower than the CPU/GPU system.

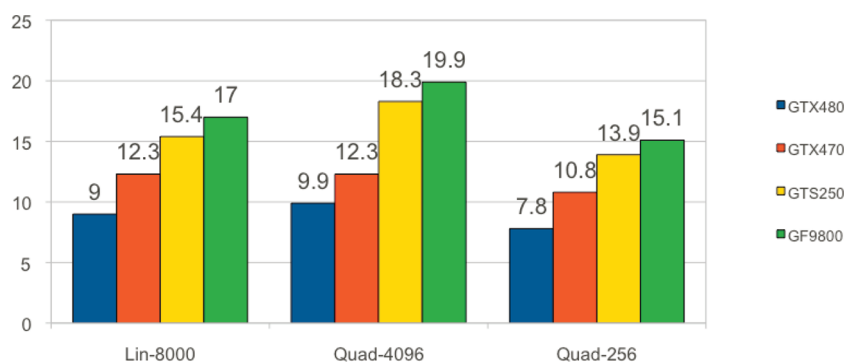


Figure 3. Timing in ms/step for calculations of nonbonded interactions for different GPUs and different options for lookup tables. The rest of the calculations in the GPU presented in this paper were done on the GTX480. The numbers below include force calculations and communication between the GPU and the CPU. Quad-256 is a table with quadratic interpolation of 256 entries that resides in the shared memory. It is the fastest option and retains acceptable accuracy (Table 2). Quad-4096 is a quadratic table with 4096 entries, which is our most accurate option and only moderately more expensive than the other options. It is too large to fit in the shared memory and resides mostly in the texture memory. Lin-8000 is a linear interpolation scheme that resides mostly in the texture memory and offers no advantage, accuracy- or speed-wise, compared to Quad-256. Coalescing the data is critical. If the data of the nonbonded list matrix M_{ij} are not coalesced, then running on GTX480 the Quad-4096 option increases the time from 9.9 to 16.5 ms.

III.4. Tryptophan Zipper (SWTWEGNKWSWK). In this solvated system, we have a total of 5847 particles, with 1875 TIP3P water molecules. The box size was $40 \times 40 \times 40 \text{ \AA}^3$, and only 16^3 mesh points were used for the reciprocal sum. The time step was 2 fs, and no RESPA integration was used. MSHAKE and SHKL were used. The cutoff distances were 9.0 and 8.5, respectively, and the nonbonded list was updated every seven steps. The system with four OpenMP threads produces 45 ns/day. Interestingly, 35.5% of the CPU/GPU calculations are conducted asynchronously. Real space nonbonded interactions require 2.2 ms/step (GPU), and generating the nonbonded list required 0.50 ms/step (GPU). On the CPU, we have the covalent forces (0.14 ms/step), the reciprocal sum (1.73 ms/step), SHAKE (0.41 ms/step), and Verlet integration (0.19 ms/step).

The one core single precision calculation yielded 3.8 ns per day in this case with the same input (2 fs time step), indicating a factor of ~ 12 speedup.

III.5. Solvated Cyclic Peptide (CA₄C) (Peptide II). This is a particularly small system of only 2690 atoms (including 882 water molecules). The system is embedded in a box of 30.1^3 \AA^3 with only 16^3 grid points for the PME calculations. The cutoffs were 10 \AA and 9 \AA . MSHAKE was used for water molecules. The time step was 1 fs, and RESPA with a second step of 4 fs was used. It was conducted in the NVE ensemble. Four threads were effective in this calculation, producing 44.1 ns/day.

The one core single precision calculations yielded in this case 3.6 ns per day. We obtained here a speedup of more than a factor of 10. This is a gratifying result, since the system is small, suggesting that our optimization is not restricted to particularly large systems.

III.6. Parameter Choices, Different Options, and Energy Conservation. In this section, we focus on the results of the DHFR system, which is a standard benchmark in the field and describe a variety of acceptable options. We provide a deeper analysis of some of our computer experiments.

III.6.1. Lookup Tables. Consider first different choices for a lookup table. The lookup table can be made very dense (a large number of points) with an interpolation scheme between the points that is simple and cheap. Indeed, NAMD^{7b} is using a

linear interpolation scheme for their lookup table and places the table in the texture memory. This is an effective choice. However, improvement can be made in efficiency and accuracy. Linear, dense tables are too large to be placed in the shared memory. Instead of a highly dense interpolation scheme, an alternative is to use significantly smaller number of table entries and use a more accurate, quadratic interpolation (see Supporting Information A and C). A sufficiently smaller table can fit into the shared memory. The disadvantage of quadratic interpolation is that it requires more calculations. However, since memory access is so crucial, additional calculations while having the data in rapidly accessed memory positively influenced the performance. Figure 3 illustrates some of these considerations

III.6.2. Energy Conservation. To better appreciate the compromises between accuracy and speed, we performed the following estimates of the energy drifts. The calculations presented below compute the reciprocal sum of PME in different grid sizes (32, 52, 64) and precisions (double and single). MSHAKE (Matrix SHAKE on water molecules) is always included. The use of SHAKE on all bonds, or light atoms, or not at all is considered. The calculations use RESPA with 4 fs for integration of the reciprocal force and 1 fs for integrating over all other forces. As mentioned earlier, a better partition of the nonbonded forces to slow and fast components is possible¹⁶ and is likely to further influence energy conservation (while requiring more computer resources). Such a study is a topic of future work. The drift in energy is estimated from a linear fit of the relative drift as a function of the time from nanosecond simulations (Figure 4).

If the drift is exceptionally small (as is the case in the most conserving options), the slope is hard to estimate, and it is ill determined. This is because it is close to zero, and even small fluctuations of the overall energy will cause significant change (with respect to zero). However, for most cases that show significant drift, the linear fit can be done accurately. We report the expected drift (using linear interpolation) on the microsecond time scale since this is the time scale we all want to be using, and the fits are indeed close to linear. The drift is expressed by the percentage of total change per microsecond (the total energy is about $-60\,000 \text{ kcal/mol}$). For comparison, we also provide the

drift reported by the authors of ACEMD in ref 8. We comment that ACEMD was reported to achieve 17.55 ms/step for one core and one GPU while we are using 38.10 ms/step. ACEMD parallelizes the reciprocal sum of the PME on the GPU. We however parallelized the PME on the CPU. Using a single core

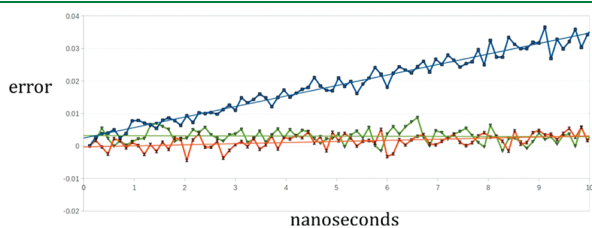


Figure 4. Energy drift for DHFR expressed as a percentage of the relative error. We plot $[100((E_t - E_0)/E_0)]$ (where E_t is the energy at time t and E_0 is the initial energy) as a function of time given in nanoseconds. Data are obtained from 10 ns simulations. Blue squares correspond to the sixth entry in Table 2 (PME 64SP, Quad-4096, NB Force add SP, SHAKE all bonds). Green triangles correspond to the first entry in Table 2 (PME 64DP, Quad-4096, NB Force add DP, SHAKE all bonds), and orange hourglasses correspond to the third entry in Table 2 (PME 64DP, Quad-4096, NB Force add SP, SHAKE all bonds). The linear regression lines (same colors) can be interpolated to 1000 ns to get the drift per microsecond (values in column 6 of Table 2). Since $|E_0|$ is $\sim 60\,000$, an increment on the y axis of 0.01 is about 6 kcal/mol. Note the significant drift when particle meshed Ewald is computed in single precision (PME 64SP). Note also the small improvement obtained when the nonbonded interactions are added on the GPU in double precision (NB Force add DP). The linear fits of the green and orange curves have significant error in their linear slopes. The error of the linear slope of the blue curve is smaller.

means that the PME in MOIL is run serially. We are therefore slower by about a factor of 2, but the energy conservation of our code as is shown in Table 2 is clearly better. Energy drifts reported for other leading programs²⁵ seem at least as high as those for ACEMD, even though a GPU was not used in the former cases.

The factors that influence energy conservation are (i) the accuracy of the lookup table, controlled by the number of entries and the level of interpolation, we consider linear interpolation with 8000 entries and quadratic interpolation with 4096 and 256 table entries; (ii) single/double precision calculations of the PME; (iii) the cutoff distance; (iv) single or double precision of the summation of real nonbonded force on the GPU; and (v) the use of SHKL versus shaking all bonds.

The two most important factors to achieve good energy conservation on the submicrosecond time scale are the accuracy of the PME reciprocal sum and the SHAKE calculations. Both must be conducted at double precision. At present, we also use double precision arithmetic in the calculation of the covalent energies and forces and in the numerical integration of the equations of motion, since the computational cost of these terms is negligible and there is no reason not to use double precision. In fact, the only term that is not conducted at double precision in our code is the calculation of real space nonbonded interactions (on the GPU). Executing the final sum of these interactions at double precision helps and brings energy conservation to the microsecond domain, but at some computational cost.

We comment that in milestone calculations of long time kinetics the trajectories used are relatively short. For example, in the calculation of the recovery stroke in myosin (a millisecond process), only subnanosecond trajectories were used.⁴ Hence, energy conservation on the nanosecond time scale may be

Table 2. Summary of Different Run Options and Their Effect on the Observed Energy Drift^a

lookup	PME	cut (box-calcd)	NB Force add	SHAKE B/L (tol 10^{-12})	DRIFT (% μ s)
Quad-4096	DP64	10.37–9.5	DP	B	0.016
Quad-4096	DP64	9.8–9.0	DP	L	0.053
*Quad-4096	DP64	10.37–9.5	SP	B	0.27
Quad-4096	DP64	8.89–8.5	SP	B	0.52
Lin-8000	DP64	10.37–9.5	SP	B	0.82
*Quad-4096	SP64	10.37–9.5	SP	B	3.2
Quad-256	SP32	8.89–8.5	SP	L	2.4
Quad-256	DP32	8.89–8.5	SP	L	0.84
**Quad-256	DP32	8.89–8.5	SP	L	−0.60
**Quad-256	SP52	8.89–8.5	SP	L	3.57
**Quad-256	DP52	8.89–8.5	SP	L	−0.30
Lin-8000	SP32	N/A-8.9	SP	NO	40.
Quad-4096	SP32	8.89–8.5	SP	NO	43.0
*Quad-4096	DP64	9.8–9.0	SP	NO	32.0
ACEMD				NO	10.15

^a A "*" in the lookup column indicates that the run was done on GTX250. The symbol "***" is for runs on the GTX470. The rest of the runs were conducted on GTX480. The energy drift is expressed as a percentage of the total change in a microsecond. Quad-4096 is a table lookup for non-bonded interactions with 4096 entries. Lin-8000 is a lookup table that is using linear interpolation. Quad-256 is a table with 256 entries that is using quadratic interpolation. The PME calculation is done either in double precision (DP) or in single precision (SP). The cut provides the two cutoff distances that are used in MOIL. The NB Force option is the use of single precision or double precision in adding the non-bonded interactions. SHAKE B/L are the options of SHAKING all bonds (B) or just light particles (L), which usually means hydrogen atoms. The tolerance of the SHAKE algorithm is fixed at 10^{-12} relative error and the reciprocal sum of PME at 10^{-9} . Note that the first row in the table is the most accurate option, and the observed drift is within the noise level. Note also the horrible drift seen with 1 fs time step and without the application of SHAKE (all water molecules are kept rigid with a Matrix SHAKE algorithm). ACEMD results are taken from reference 8 for a 1 fs time step. Note that we tried two grid sizes for the reciprocal sum calculations, grid sizes of 32 and 64 points at each Cartesian direction. The effect on energy conservation was very small. Grid size impacts, however, pair correlation functions, and an appropriate choice must be made for the specific system at hand. See the text for more details.

sufficient for the theory-based long time dynamics and kinetic calculations of MOIL.

IV. CONCLUDING REMARKS

We describe an implementation of MOIL-opt, a version of the dynamics module of the program MOIL¹¹ that was ported to a high performance laboratory tool, namely, a CPU with several cores (we used up to four) and a GPU card. The new features of the present version include a novel implementation of a lookup table, careful partitioning of the work between the GPU and a shared memory multicore system, and a detailed analysis of the energy conservation of the simulation. It is shown that the prime factor for successful energy conservation is the calculations of the reciprocal space of the Ewald summation at double precision. The double precision summation of the nonbonded interactions on the GPU comes next.

MOIL strongly emphasizes the calculations of reaction mechanisms and kinetics. It provides a set of tools to compute reaction paths and approximate long time trajectories and more recently also the tools of milestoning. Using a midrange graphic card, a system can be built that provides both high-end performance and accurate energies. Sampling correctly from the microcanonical ensemble and the production of energy conserving trajectories is necessary for estimating microscopic time scales and rates. Others used stochastic dynamics to overcome the energy drift. However, it is not clear if stochastic dynamics, which employs a basic integrator that does not conserve energy even at the zero friction limit, produces configurations with probability $p(x)$ proportional to $\exp(-U(x)/kT)$. This is in contrast to a Monte Carlo algorithm in which by rejections and acceptances of steps the Boltzmann distribution is enforced.

We emphasize that we do not object (of course) to stochastic dynamics obtaining sampling from ensembles other than microcanonical if the differential equations are solved exactly. However, information about microscopic kinetics is better obtained from microcanonical calculations. Even if the canonical ensemble is obtained, the addition of stochastic forces can significantly affect the time scale of the processes we study. For example, in the strong friction limit, the rate constant is inversely proportional to the friction coefficient of the Langevin equation. The friction coefficient is typically determined as a phenomenological (not microscopic) parameter.

The program MOIL including the recent addition to it, MOIL-opt, is freely available from <http://clsb.ices.utexas.edu/prebuilt/>.

■ ASSOCIATED CONTENT

S Supporting Information. An outline of the procedure used for quadratic interpolation of lookup tables, pseudo code for the generation of a nonbonded list for the GPU, pseudocode for real space calculations of nonbonded interactions, and a performance profile of a serial code of MOIL. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ron@ices.utexas.edu.

■ ACKNOWLEDGMENT

This research was supported by NIH grant GM059796 and NSF grant CCF-0833162 to R.E. Helpful discussions and the advice of Professor Martin Burtscher are gratefully acknowledged. We thank Mauro Lorenzo Mugnai for his help in setting

up the membrane system and Ignacio F. Gallardo for his help in setting up the simulation of peptide I.

■ REFERENCES

- (1) Olender, R.; Elber, R. Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *J. Chem. Phys.* **1996**, *105* (20), 9299–9315.
- (2) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. Transition path sampling. In *Adv. Chem. Phys.*; John Wiley & Sons Inc: New York, 2002; Vol. 123, pp 1–78.
- (3) (a) Faradjian, A. K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120* (23), 10880–10889. (b) Majek, P.; Elber, R. Milestoning without a Reaction Coordinate. *J. Chem. Theory Comput.* **2010**, *6* (6), 1805–1817. (c) Moroni, D.; Bolhuis, P. G.; van Erp, T. S. Rate constants for diffusive processes by partial path sampling. *J. Chem. Phys.* **2004**, *120* (9), 4055–4065. (d) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **2003**, *118* (17), 7762–7774. (e) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Forward flux sampling-type schemes for simulating rare events: Efficiency analysis. *J. Chem. Phys.* **2006**, *124* (19), 17.
- (4) Elber, R.; West, A. Atomically detailed simulation of the recovery stroke in myosin by Milestoning. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 5001–5005.
- (5) Darden, T.; York, D.; Pedersen, L. Particle mesh ewald - an $n \log(n)$ method for ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.
- (6) Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular-dynamics. *J. Chem. Phys.* **1992**, *97* (3), 1990–2001.
- (7) (a) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330* (6002), 341–346. (b) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. Accelerating molecular modeling applications with graphics processors. *J. Comput. Chem.* **2007**, *28* (16), 2618–2640.
- (8) Harvey, M. J.; Giupponi, G.; De Fabritiis, G. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, *5* (6), 1632–1639.
- (9) Ravikant, D. V. S.; Elber, R. PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins: Struct., Funct., Bioinf.* **2010**, *78* (2), 400–419.
- (10) Anderson, J. A.; Lorenz, C. D.; Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **2008**, *227* (10), 5342–5359.
- (11) Elber, R.; Roitberg, A.; Simmerling, C.; Goldstein, R.; Li, H. Y.; Verkhivker, G.; Keasar, C.; Zhang, J.; Ulitsky, A. Moil a program for simulations of macromolecules. *Comput. Phys. Commun.* **1995**, *91* (1–3), 159–189.
- (12) Elber, R.; Ghosh, A.; Cardenas, A. Long time dynamics of complex systems. *Acc. Chem. Res.* **2002**, *35* (6), 396–403.
- (13) (a) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of cartesian equations of motion of a system with constraints - molecular dynamics of N-alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. (b) Weinbach, Y.; Elber, R. Revisiting and parallelizing SHAKE. *J. Comput. Phys.* **2005**, *209* (1), 193–206.
- (14) van Meel, J. A.; Arnold, A.; Frenkel, D.; Zwart, S. F. P.; Belleman, R. G. Harvesting graphics power for MD simulations. *Mol. Simul.* **2008**, *34* (3), 259–266.
- (15) Harvey, M. J.; De Fabritiis, G. An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware. *J. Chem. Theory Comput.* **2009**, *5* (9), 2371–2377.
- (16) Morrone, J. A.; Zhou, R. H.; Berne, B. J. Molecular Dynamics with Multiple Time Scales: How to Avoid Pitfalls. *J. Chem. Theory Comput.* **2010**, *6* (6), 1798–1804.
- (17) Schnieders, M. J.; Fenn, T. D.; Pande, V. S. Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for

Macromolecular Crystals. *J. Chem. Theory Comput.* **2011**, *7* (4), 1141–1156.

(18) Yip, V.; Elber, R. Calculation of a list of neighbors in molecular-dynamics simulations. *J. Comput. Chem.* **1989**, *10* (7), 921–927.

(19) Eastman, P.; Pande, V. S. Efficient Nonbonded Interactions for Molecular Dynamics on a Graphics Processing Unit. *J. Comput. Chem.* **2010**, *31* (6), 1268–1272.

(20) Kaminski, G.; Friesner, R.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.

(21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(22) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269–6271.

(23) (a) Jorgensen, W. L.; Tiradorives, J. The OPLS potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. (b) Berger, O.; Edholm, O.; Jahnig, F. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophys. J.* **1997**, *72* (5), 2002–2013.

(24) Gallardo, I. F.; Webb, L. J. Tethering Hydrophobic Peptides to Functionalized Self-Assembled Mono layers on Gold through Two Chemical Linkers Using the Huisgen Cycloaddition. *Langmuir* **2010**, *26* (24), 18959–18966.

(25) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GRO-MACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

Size-Dependent Phase Changes in Water Clusters

Toshihiro Kaneko,[†] Takuma Akimoto,[†] Kenji Yasuoka,^{*,†} Ayori Mitsutake,[‡] and Xiao Cheng Zeng^{*,§}[†]Department of Mechanical Engineering and [‡]Department of Physics, Keio University, Yokohama, 223-8522, Japan[§]Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588, United States

ABSTRACT: We investigate melting behavior of water clusters $(\text{H}_2\text{O})_N$ ($N = 7, 8, 11,$ and 12) by using multicanonical-ensemble molecular dynamics simulations. Our simulations show that the melting behavior of water clusters is highly size dependent. Based on the computed canonical average of the potential energy and heat capacity C_V , we conclude that $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$ exhibit first-order-like phase change, while $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$ exhibit continuous-like phase change. The melting temperature range for $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$ can be defined based on the peak position of $C_V(T)$ and $dC_V(T)/dT$ (where T is the temperature). Moreover, for $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$, the solid- and liquid-like phases separate temporally in the course of simulation. In contrast, no temporal separation of solid- and liquid-like phases is observed for $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$. In light of the notable temporal separation of solid- and liquid-like phases for $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$, an alternative computer approach for estimating the melting temperature range is proposed based on the time-dependent Lindemann parameters. We find that the melting temperature range estimated from both definitions is consistent with each other for $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$ but not for $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$. We also find that the melting behavior of small water clusters can be conveniently assessed if the energy differences of neighbor-sized clusters at zero temperature are known.

1. INTRODUCTION

Freezing of bulk water is a classical example of the first-order phase transition. Even in highly confined systems, e.g., when water is confined to a carbon nanotube or between two graphene sheets with a nanoscale separation, the freezing transition is still the first order but can become continuous under certain conditions due to the confinement effects.^{1,2} On the other hand, for finite-size systems, such as molecular or atomic clusters, the melting-like phase change can be dramatically different from their bulk counterparts. The first experimental evidence of size-dependent melting behavior in atomic clusters was reported by Schmidt et al.³ More recently, Jarrold and co-workers have observed that small gallium and tin clusters melt at temperatures higher than the melting temperature of bulk metals.^{4,5} Although the melting behavior of clusters is strongly size dependent, it would be desirable to seek some generic features on their phase-change behavior.

Small water clusters have attracted considerable interest because the underlying intermolecular interactions and hydrogen-bonding dynamics can be more precisely described on the molecular level in these finite-size systems. In turn, the study of the finite-size water clusters provides more accurate models of water, which may lead to improved understanding of bulk water.^{6–24}

Unlike bulk matters, it has been shown that different “phases” can coexist in clusters dynamically without showing phase separation.^{25,26} Indeed, dynamical coexistence and reversible structural change have already been observed in simulations of $(\text{H}_2\text{O})_6$, $(\text{H}_2\text{O})_8$, and $(\text{H}_2\text{O})_{20}$ water clusters by a number of research groups.^{6,15,16,27} Tsai and Jordan⁶ perhaps were the first to show a simulation evidence of the solid-to-liquid like transition in $(\text{H}_2\text{O})_8$ clusters. Later, Laria et al.¹⁵ demonstrated that the transition between liquid and solid $(\text{H}_2\text{O})_8$ did not involve the passage over an energetic barrier. Shin et al.¹⁸ and Frantsuzov et al.²² have investigated quantum effects on the melting of $(\text{H}_2\text{O})_8$ cluster. Both groups have shown that the quantum effects shift the melting temperature toward a lower value, about 8–25 K below the value

obtained from the classical simulation for $(\text{H}_2\text{O})_8$. On the other hand, for small-sized water clusters with odd number of molecules, such as $(\text{H}_2\text{O})_7$, Tharrington and Jordan¹⁶ have shown that they typically do not undergo a well-defined melting transition; in particular, $(\text{H}_2\text{O})_7$ may display “glass-like” behavior. Today, it has been well established that the dynamical coexistence in small clusters is due to the existence of many local minima on the potential energy surface, and as such, it can be difficult to exactly compute the heat capacity from conventional molecular dynamics simulation. Since the heat capacity is a very important parameter to characterize phase transition, in this study, we have employed an efficient simulation method, namely, the multicanonical (MUCA)-ensemble method to compute the canonical average of heat capacity. The MUCA-ensemble method was originally developed by Berg and Neuhaus^{28,29} and later extended to work with molecular dynamics (MD) simulation methods.^{30,31} Compared to conventional MD methods, the MUCA-MD can sample a wider range of potential energy surface without having the system trapped in local minima. In a previous study, we investigated melting behavior of the Lennard-Jones clusters using the MUCA-MD method.³² Here, based on the MUCA-MD simulations, we show that small water clusters entail much richer melting behaviors than small LJ clusters.

2. MODEL AND SIMULATION METHOD

The TIP4P water model³³ is used in all simulations, and the free-boundary condition is applied in the MD simulations. Initial coordinates for molecules in water clusters are taken from the Cambridge Cluster Database³⁴ which provides the global minima of the clusters at 0 K. Initial velocities of molecules are set such that the momentum, and the angular momentum of the clusters are zero. To integrate Newton's equations, the velocity Verlet algorithm with SHAKE/RATTLE method is adopted.

Received: July 1, 2011

Published: August 30, 2011

The MD simulations are performed in both the MUCA and conventional canonical ensembles. Nosé–Hoover thermostat is applied to control the temperature of the systems. In the simulations, if the distance between the center of mass of molecules becomes greater than 30 Å, the system is viewed as under evaporation, and the simulation will be restarted with the same initial positions for the molecules but with different initial velocities. We have also examined an alternative criterion for defining a water cluster, namely, if the O–O distance is within 4 Å, the two relevant water molecules are viewed as being in the same cluster. A test simulation confirms that the two cluster criteria are consistent with one another.

For each size of clusters, a number of simulations with different (given) multicanonical weight factors are performed. At the end of simulations, results for cluster of the same size are combined by using the weighted histogram analysis method^{35–37} to calculate the density of state $n(E)$, where E is potential energy. The simulations are repeated until $n(E)$ is converged. Once $n(E)$ is known, the canonical average of a physical observable $A(E)$ at given temperature T is given by

$$\langle A \rangle_T = \frac{\sum_E A(E)n(E) \exp(-E/k_B T)}{\sum_E n(E) \exp(-E/k_B T)} \quad (1)$$

Based on eq 1, the average potential energy $\langle E \rangle_T$ can be calculated. To estimate the melting temperature of $(\text{H}_2\text{O})_N$ clusters, we calculate the heat capacity at constant volume C_V and the temperature derivative dC_V/dT based on the following equations:

$$C_V = \frac{\langle E^2 \rangle_T - \langle E \rangle_T^2}{k_B T^2} + 3Nk_B \quad (2)$$

$$\frac{dC_V}{dT} = \frac{\langle (E - \langle E \rangle_T)^3 \rangle_T}{k_B^2 T^4} - \frac{2\langle (E - \langle E \rangle_T)^2 \rangle_T}{k_B T^3} \quad (3)$$

where $\langle E \rangle_T$, $\langle E^2 \rangle_T$, and $\langle E^3 \rangle_T$ are computed using eq 1.

3. RESULTS AND DISCUSSION

3.1. Static Properties. The computed canonical average of potential energy E , heat capacity C_V and the temperature derivative dC_V/dT are summarized in Figure 1. Two distinct types of curves can be seen in Figure 1: For $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$, the peaks of C_V can be easily identified at 181.6 and 164.6 K, respectively, in Figure 1b; and the corresponding S-shaped curves of dC_V/dT can be observed in Figure 1c. Similar behavior of the C_V – T curve has been reported previously by Tharrington and Jordan.¹⁶ For $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$, however, no apparent peaks are seen in the C_V – T curves. Instead, modest peaks in the dC_V/dT curves are observed at 122.5 and 163.6 K, respectively in Figure 1c. There is no clear change of slope in the potential energy curves in Figure 1a. According to the Ehrenfest classification, the first-order phase transition exhibits a discontinuity (like a step function) in the first derivative of the free energy, such as the potential energy, so that the heat capacity curve should exhibit a very sharp peak, like a δ function. The second-order continuous phase transition is continuous in the first derivative of the free or potential energies but exhibits a discontinuity in the second derivative of the free energy, such as the heat capacity. For finite-size cluster systems, the sharp discontinuity is smeared out due to the finite-size effect, resulting in more broad peaks for C_V or dC_V/dT curves.³⁸ Thus, based on calculation of the canonical averaged E ,

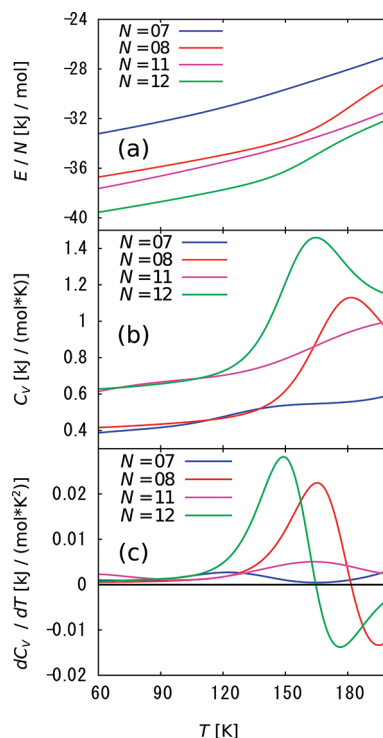


Figure 1. The canonical average of (a) potential energy E , (b) heat capacity C_V , and (c) the temperature derivative dC_V/dT .

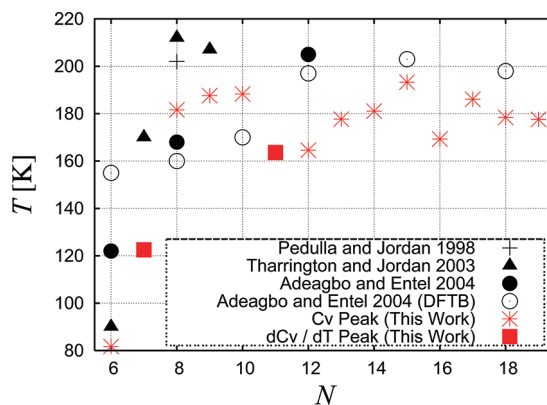


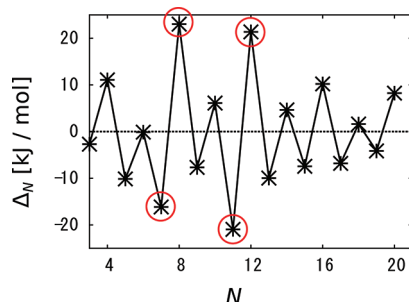
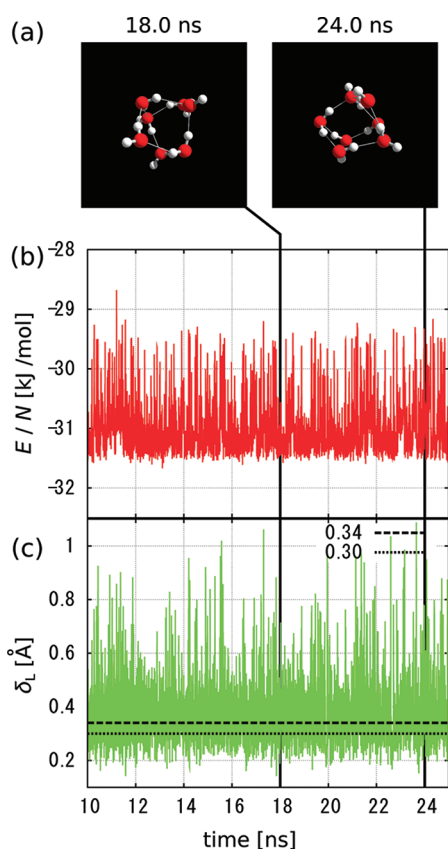
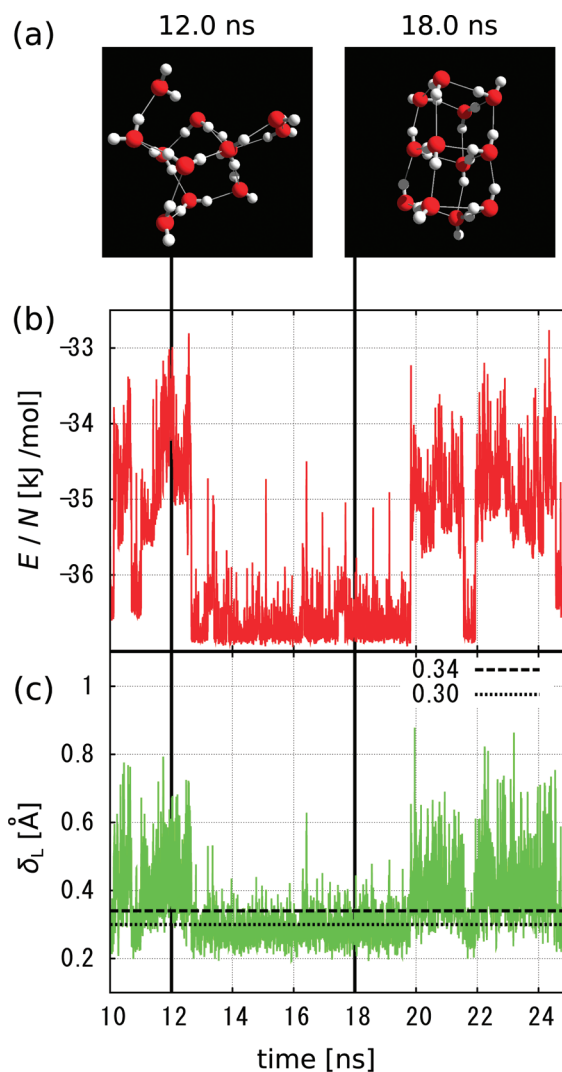
Figure 2. Estimated melting points for each size of small water clusters. For $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$, the melting points are defined by the peak position of dC_V/dT , while for other sizes, the melting points are defined by the position of C_V peak. The water model is TIP4P, except “Adeagbo and Entel, 2004 (density functional-based tight binding)”.

C_V and dC_V/dT , we suggest that the first-order-like melting transition occurs for $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$ and continuous-like melting transition occurs for $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$.

Due to the finite-size effects, the melting transition of small clusters does not occur at a single temperature but within a temperature range.³⁸ Hence, definitions of melting temperature are no longer unique. One possible definition (or C_V based definition) for the melting temperature range associated the first-order-like phase change can be that between the position of the dC_V/dT peak and the position of C_V peak. For clusters exhibiting continuous-like melting transition, such as $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$, one arbitrary C_V -based definition for the melting temperature could be the position of the dC_V/dT peak. In previous

Table 1. Summary of Estimated Melting Temperature Ranges

N	C_V peak, K	dC_V/dT peak, K	Lindemann parameter, K
7	—	122.5	100–120
8	181.6	165.3	165–180
11	—	163.6	110–130
12	164.6	149.0	150–160

**Figure 3.** Calculated Δ_{N_i} based on eq 4, for each cluster. It has the maximum values at $N = 8$ and 12 , and the corresponding global-minimum structures are stable. It has the minimum values at $N = 7$ and 11 , and the corresponding global-minimum structures are unstable.**Figure 4.** (a) Snapshots of typical structures at two different times, one corresponding to relatively high-energy state and another corresponding to low-energy state (b) potential energy evolution $E(t)$, and (c) the Lindemann parameter $\delta_L(t)$ for $(\text{H}_2\text{O})_7$ at 120 K. $E(t)$ and $\delta_L(t)$ are averaged value over every 2.5 ps. No apparent two states (high- and low-energy states) are observed.**Figure 5.** (a) Snapshots of typical structures at two different time, one corresponding to high-energy (liquid-like) state (phase) and another corresponding to low-energy (solid-like) state, (b) potential energy evolution $E(t)$, and (c) the Lindemann parameter $\delta_L(t)$ for $(\text{H}_2\text{O})_{12}$ at 150 K. $E(t)$ and $\delta_L(t)$ are averaged value over every 2.5 ps. A temporal intermittency between two states (phases) is clearly seen in (b) and (c). Two horizontal dashed and dotted lines in (c) mark a threshold range for $\delta_{L,\text{th}}$ defined to be within 0.30–0.34 Å.

simulations, Pedulla and Jordan calculated C_V of $(\text{H}_2\text{O})_6$ and $(\text{H}_2\text{O})_8$ using Monte Carlo simulation.¹¹ Later, Wales and Ohmine⁷ revealed the alternation of liquid- and solid-like behavior of TIP4P $(\text{H}_2\text{O})_8$ cluster in the course of computer simulation. Their estimated melting temperature was about 190 K. Rodriguez et al.¹² monitored phase transition behavior of water clusters by following the temperature dependence in dipole moments, number of hydrogen bonds, and Lindemann's parameter. They found the melting temperature of TIP4P $(\text{H}_2\text{O})_8$ cluster was about 160 K. Tharrington and Jordan also estimated the melting temperature of TIP4P $(\text{H}_2\text{O})_{n=6-9}$ based on the position of the C_V peak using the parallel-tempering Monte Carlo method.¹⁶ They found the melting temperature of TIP4P $(\text{H}_2\text{O})_8$ cluster was about 215 K, higher than those reported previously. Note that Tharrington and Jordan used a constraining sphere (with radius 4.25 Å) to prevent evaporation. This constraining sphere may incur some effective

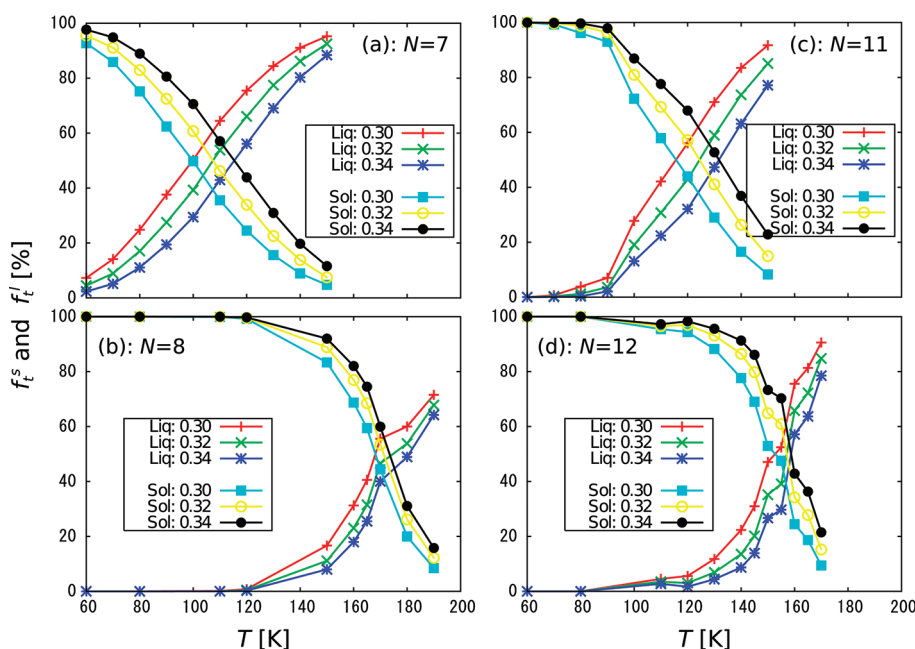


Figure 6. Fraction of time for the system being in the liquid- or solid-like state, determined based the threshold value $\delta_{L,\text{th}} = 0.30, 0.32, \text{ and } 0.34$, respectively. If $\delta_L(t) < \delta_{L,\text{th}}$, then the cluster is viewed to be in the solid-like state, and if $\delta_L(t) > \delta_{L,\text{th}}$, then the cluster is viewed to be in the liquid-like state.

“pressure” on the $(\text{H}_2\text{O})_8$ cluster, thereby leading to a higher melting temperature for the TIP4P $(\text{H}_2\text{O})_8$ cluster. Adeagbo and Entel estimated the melting temperature based on the Lindemann parameter computed from both classical MD and density functional tight-binding (DFTB) simulations.¹⁷ Despite the estimated melting temperature of TIP4P $(\text{H}_2\text{O})_8$ cluster ranges from 160 to 215 K among different studies (see Figure 2), the qualitative features of the C_V – T curves from all simulations are the same. The estimated melting temperature ranges or melting temperature are summarized in Table 1.

To gain more insights into the size-dependent phase change, we compute the energy difference Δ_N of neighbor-sized clusters, given by the following equation:

$$\Delta_N = E_{\min}(N + 1) + E_{\min}(N - 1) - 2E_{\min}(N) \quad (4)$$

where $E_{\min}(N)$ is the potential energy of the global minimum at 0 K published previously.^{34,39,40} Δ_N corresponds to $d^2E_{\min}(N)/dN^2$ or the curvature of $E_{\min}(N)$. The results are plotted in Figure 3, where the four clusters considered in this study are marked by red circles. Clearly, Δ_N shows the highest values for $N = 8, 12$ but the lowest values for $N = 7, 11$. We suggest that when Δ_N is greater than 20 kJ/mol, like $(\text{H}_2\text{O})_8$ and $(\text{H}_2\text{O})_{12}$, the corresponding global-minimum structure is very stable, and once the system is trapped in this structure it is difficult to see a phase change for the system. On the other hand, when Δ_N reaches a large negative value (beyond -15), like $(\text{H}_2\text{O})_7$ and $(\text{H}_2\text{O})_{11}$, the corresponding global-minimum structure is relatively unstable, and the system cannot be trapped in this structure. Hence, the magnitude of Δ_N value offers a guide to the melting behavior (first-order- or continuous-like) for small water clusters. In addition, we have attempted to examine whether the melting temperature of water clusters would converge at certain size. To this end, we have performed several independent simulations to estimate melting temperature of water clusters with size $N = 6, 9, 10, 13\text{--}19$. Interestingly, we find that the melting temperature of $(\text{H}_2\text{O})_N$ clusters seems to converge towards 180 K as the size increases (see Figure 2).

3.2. Time-Dependent Properties. To gain additional insights into the size-dependent melting behavior, we further analyze the time-dependent potential energy $E(t)$ and the Lindemann parameter $\delta_L(t)$ of the system in the canonical ensemble. To this end, either 100.05 or 80.05 ns total simulation time is used, in which the initial 0.05 ns is for relaxing the system and is thus excluded from the analysis. The Lindemann parameter $\delta_L(t)$ is given by the equation:

$$\delta_L(t) = \sqrt{\frac{1}{N} \sum_{i=1}^N [\langle r_{\text{O}_i}^2(t) \rangle_\tau - \langle r_{\text{O}_i}(t) \rangle_\tau^2]} \quad (5)$$

where N is the number of particles in the cluster, $r_{\text{O}_i}(t)$ is the oxygen position of molecule i , and $\langle \dots \rangle_\tau$ denotes the time average over a time interval τ (in this study, $\tau = 2.5$ ps). $E(t)$ is also averaged over 2.5 ps time interval.

Results of $(\text{H}_2\text{O})_7$ at 120 K are shown in Figure 4. The two snapshots shown in Figure 4a, one corresponding to relatively high-energy state at $t = 18.0$ ns and another corresponding to low-energy state at $t = 24.0$ ns, do not show notable difference, suggesting that $(\text{H}_2\text{O})_7$ can easily change its conformation structure. Hence, the melting behavior of $(\text{H}_2\text{O})_7$ is difficult to discern, which is also reflected by the feature-less behavior of $E(t)$ and $\delta_L(t)$. Similar melting behavior is observed for $(\text{H}_2\text{O})_{11}$. Results of $(\text{H}_2\text{O})_{12}$ at 150 K are shown in Figure 5. Unlike $(\text{H}_2\text{O})_7$, a temporal intermittency between two states is clearly seen, one state is the high-energy (liquid-like) state (phase) and the other is the low-energy (solid-like) state. This means that the cluster exhibits phase change back and forth in the course of simulation, as shown by two typical snapshots in the system, one at $t = 12.0$ ns and another at $t = 18.0$ ns (Figure 5a). Note that this temporal separation of low- and high-energy structure in $(\text{H}_2\text{O})_{12}$ was also observed in $(\text{H}_2\text{O})_{20}$ at 120 K by Nishio and Mikami.²⁷ Moreover, the time-dependent Lindemann parameter $\delta_L(t)$ shown in Figure 5c suggests that the system stays in the low-energy solid-like state longer than in the high-energy liquid-like state. Based on this observation, we assign a threshold region for $\delta_L(t)$, that is, 0.30–0.34 Å. Note that the

lattice constant of bulk ice is about 3 Å. Hence, 10% of the lattice constant of bulk ice is a sensible parameter to distinguish liquid and solid phases. Using threshold values $\delta_{L,th} = 0.30, 0.32,$ and $0.34,$ respectively, the fraction of time for the system in the solid- and liquid-like states can be estimated. If $\delta_L(t) < \delta_{L,th}$, then the cluster is viewed as in the solid-like state, and if $\delta_L(t) > \delta_{L,th}$, then the cluster is viewed as in the liquid-like state. The results at various temperatures are summarized in Figure 6. Based on the estimated fraction of time f_t^s and f_t^l for the system being in solid- and liquid-like states, we propose another possible definition for the melting temperature range, that is, the temperature range between the crossing points ($f_t^s = f_t^l = 0.5$) for $\delta_{L,thr} = 0.3$ and 0.34 . Newly estimated melting temperature ranges are given in Table 1. For $(H_2O)_8$ and $(H_2O)_{12}$, we find that the newly estimated melting temperature range is consistent with that based on the position of C_V and dC_V/dT peak. However, for $(H_2O)_7$ and $(H_2O)_{11}$, the newly estimated melting temperature range is notably lower than the melting temperature estimated based on the position of dC_V/dT peak. Hence, the latter may be viewed as an upper limit for the melting temperature of small water clusters that shows the second-order-like phase change.

4. CONCLUSION

We have employed the multicanonical-ensemble molecular dynamics simulation method to investigate melting behavior of small water clusters $(H_2O)_N$ ($N = 7, 8, 11, 12$). Our simulations confirm that the melting behavior of small water clusters is highly size dependent. In particular, $(H_2O)_8$ and $(H_2O)_{12}$ exhibit first-order-like phase change, while $(H_2O)_7$ and $(H_2O)_{11}$ exhibit continuous-like phase change. Moreover, for $(H_2O)_8$ and $(H_2O)_{12}$, the solid- and liquid-like phases separate temporally in the course of simulation. In contrast, no temporal separation of solid- and liquid-like phases is observed for $(H_2O)_7$ and $(H_2O)_{11}$.

We have proposed two definitions for estimating the melting temperature range associated with the first-order-like phase change: one based on the peak position of $C_V(T)$ and dC_V/dT and another based on the time-dependent Lindemann parameters (in view of notable temporal separation of solid- and liquid-like phases in these clusters). We find that the melting temperature range estimated from both definitions are consistent with each other for $(H_2O)_8$ and $(H_2O)_{12}$. Finally, we suggest that the melting behavior of small water clusters can be conveniently assessed if the energy differences Δ_N of neighbor-sized clusters at zero temperature are known.

AUTHOR INFORMATION

Corresponding Author

*E-mail: yasuoaka@mech.keio.ac.jp; xczeng@phase2.unl.edu.

ACKNOWLEDGMENT

T.K. was supported by a Grant in Aid for Japan Society for the Promotion of Science (JSPS) Fellows 23-2003 of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) in Japan. X.C.Z. was supported by U.S. NSF (grant nos. EPS-1010674 and CBET-1066947) and A.R.L. (grant no. W911NF1020099).

REFERENCES

- (1) Koga, K.; Gao, G. T.; Tanaka, H.; Zeng, X. C. *Nature (London)* **2001**, *412*, 802–805.
- (2) Han, S.; Choi, M. Y.; Kumar, P.; Stanley, H. E. *Nat. Phys.* **2010**, *6*, 685–689.

- (3) Schmidt, M.; Kusche, R.; von Issendorff, B.; Haberland, H. *Nature (London)* **1998**, *393*, 238–240.
- (4) Breaux, G. A.; Cao, B.; Jarrold, M. F. *J. Phys. Chem. B* **2005**, *109*, 16575–16578.
- (5) Breaux, G. A.; Neal, C. M.; Cao, B.; Jarrold, M. F. *Phys. Rev. B* **2005**, *71*, 073410.
- (6) Tsai, C. J.; Jordan, K. D. *J. Chem. Phys.* **1991**, *95*, 3850–3853.
- (7) Wales, D. J.; Ohmine, I. *J. Chem. Phys.* **1993**, *98*, 7257–7268.
- (8) Gregory, J. K.; Clary, D. C. *J. Phys. Chem.* **1996**, *100*, 18014–18022.
- (9) Cruzan, J. D.; Braly, L. B.; Liu, K.; Brown, M. G.; Loeser, J. G.; Saykally, R. J. *Science* **1996**, *271*, 59–62.
- (10) Liu, K.; Brown, M. G.; Carter, C.; Saykally, R. J.; Gregory, J. K.; Clary, D. C. *Nature* **1996**, *381*, 501–503.
- (11) Pedulla, J. M.; Jordan, K. D. *J. Chem. Phys.* **1998**, *239*, 593–601.
- (12) Rodriguez, J.; Laria, D.; Marceca, E.; Estrin, D. *J. Chem. Phys.* **1999**, *110*, 9039–9047.
- (13) Nauta, K.; Miller, R. E. *Science* **2000**, *287*, 293–295.
- (14) Keutsch, F. N.; Saykally, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10533–10540.
- (15) Laria, D.; Rodriguez, J.; Dellago, C.; Chandler, D. *J. Phys. Chem. A* **2001**, *105*, 2646–2651.
- (16) Tharrington, A. N.; Jordan, K. D. *J. Phys. Chem. A* **2003**, *107*, 7380–7389.
- (17) Adeagbo, W. A.; Entel, P. *Phase Transitions* **2004**, *77*, 63–79.
- (18) Shin, S.; Son, W.; Jang, S. *J. Mol. Struct. (Theochem)* **2004**, *673*, 109–113.
- (19) Aguado, A.; López, J. M. *Phys. Rev. Lett.* **2005**, *94*, 233401.
- (20) Bulusu, S.; Yoo, S.; Aprà, E.; Xantheas, S.; Zeng, X. C. *J. Phys. Chem. A* **2006**, *110*, 11781–11784.
- (21) Langley, S. F.; Curotto, E.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **2007**, *126*, 084506.
- (22) Frantsuzov, P. A.; Mandelshtam, V. A. *J. Chem. Phys.* **2008**, *128*, 094304.
- (23) Hock, C.; Schmidt, M.; Kuhnen, R.; Bartels, C.; Ma, L.; Haberland, H.; v.Issendorff, B. *Phys. Rev. Lett.* **2009**, *103*, 073401.
- (24) Yoo, S.; Apra, E.; Zeng, X. C.; Xantheas, S. *J. Phys. Chem. Lett.* **2010**, *1*, 3122–3127.
- (25) Labastie, P.; Whetten, R. L. *Phys. Rev. Lett.* **1990**, *65*, 1567–1570.
- (26) Wales, D. J.; Berry, R. S. *Phys. Rev. Lett.* **1994**, *73*, 2875–2878.
- (27) Nishio, K.; Mikami, M. *J. Chem. Phys.* **2009**, *130*, 154302.
- (28) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249–253.
- (29) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9–12.
- (30) Hansmann, U. H. E.; Okamoto, Y.; Eisenmenger, F. *Chem. Phys. Lett.* **1996**, *259*, 321–330.
- (31) Nakajima, N.; Nakamura, H.; Kidera, A. *J. Phys. Chem. B* **1997**, *101*, 817–824.
- (32) Kaneko, T.; Yasuoka, K.; Mitsutake, A.; Zeng, X. C. Proceedings of the ASME/JSME 2011 8th Thermal Engineering Joint Conference, Honolulu, Hawaii, March 13–17, 2011; ASME: New York, 2011; AJTEC2011–44457.
- (33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *76*, 926–935.
- (34) Wales, D. J.; Doye, J. P. K.; Dullweber, A.; Hodges, M. P.; Naumkin, F. Y.; Calvo, F.; J. Hernández-Rojas Middleton, T. F. *The Cambridge Cluster Database*; Department of Chemistry, Cambridge University: Cambridge, U.K.; <http://www-wales.ch.cam.ac.uk/CCD.html>.
- (35) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.
- (36) Kumar, S.; Bouzida, D.; Swedensen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (37) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *J. Chem. Phys.* **2003**, *118*, 6664–6675.
- (38) Berry, R. S.; Beck, T. L.; Davis, H. L.; Jellinek, J. *Adv. Chem. Phys.* **1988**, *70* (2), 75–138.
- (39) Wales, D. J.; Hodges, M. P. *Chem. Phys. Lett.* **1998**, *286*, 65–72.
- (40) Kazachenko, S.; Thakkar, A. J. *Chem. Phys. Lett.* **2009**, *476*, 120–124.

Multireference Double Electron Attached Coupled Cluster Method with Full Inclusion of the Connected Triple Excitations: MR-DA-CCSDT

Monika Musiał,^{*,†} Stanisław A. Kucharski,[†] and Rodney J. Bartlett[‡]

[†]Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland

[‡]Quantum Theory Project, Departments of Chemistry and Physics, University of Florida, Gainesville, Florida 32611, United States

ABSTRACT: The multireference (MR) double electron attached (DA) coupled cluster (CC) method with full inclusion of the connected triple excitations has been applied to study various kinds of MR situations. The MR-DA-CCSDT (S, Singles; D, Doubles; T, Triples) equations have been derived and implemented in an efficient way with n^6 scaling for the target multireference states. They can be used for producing potential energy curves (PECs) for some classes of molecules, e.g., when double molecular cations separate into two closed shell fragments, illustrated with the example of the Na₂ molecule. Correct PECs have also been obtained on dissociation of the N–N and C–C bonds in N₂H₄ and C₂H₆ molecules, respectively. Another application is the behavior of the molecular energy when we change dihedral angle in the ethylene molecule: with the MR-DA-CC, we see immediate improvement of the results with smooth, cusp free, behavior around the 90° region.

INTRODUCTION

The equation-of-motion coupled cluster (EOM-CC)^{1–12} method has been shown to be a very useful tool in the study of excited, ionized, and electron attached states with many applications.^{13–17} However, there is no systematic treatment of the double ionized or doubly electron attached states. The EOM-CC formalism developed for the double ionization (DI) and double electron attachment (DA) issues is no more complicated than those derived for the ionization potential (IP) and electron attachment (EA) or excitation energy (EE) problems and can even have lower computational scaling. But in order to achieve tractable equations, we need to take advantage of the various factorization procedures which formally complicate the equations but significantly reduce the computational effort.

Another reason why the DA and DI problems have attracted less attention is that their experimental verification is more difficult. However, both DI and DA approaches can be successfully used—as we are going to show in this work—to evaluate other molecular properties than direct DA values. The terminology multireference (MR) DA-CC pertains to approximations that can be made in R that are similar to those in MR configuration interaction (CI) and found to be quite important but lie outside the standard approximations anticipated in EOM-CC. The first is to have different levels of excitation operators in R and in T ,¹⁸ and this, in particular, provides a MR-DA-CCSD (S, Singles; D, Doubles) solution that is virtually identical to MR-DA-CCSDT (T, Triples), at $\sim n^2$ less cost.

In previous work, we developed the EOM-CC scheme for DI calculations (DI-EOM-CC=MR-DI-CC)¹⁹ (see also similar theoretical works^{20,21}). The double ionization potential quantities^{20–27} are in many cases experimentally available, for example, via Auger spectroscopy, double charge transfer spectroscopy (DCT), and/or threshold photoelectron coincidence spectroscopy (TPEsCO). Much less attention has been paid to the DA topic since the latter represents a more complex problem from the experimental viewpoint, and also the theoretical apparatus used to extract

the required quantities from the experimental data is much more involved.^{22,28–30} However, DA calculations, including some earlier work based on the similarity transformed EOM (STEOM)^{22,28–30} limited to CCSD,^{22,28} are suited to generating potential surfaces or for calculating excitation energies for some systems with pronounced multireference character. The determination of excitation energies via the double electron attachment scheme also corresponds to obtaining wave functions as proper spin-eigenfunctions provided the doubly positive ion is closed shell. This feature is extremely important for open-shell systems, as even normal single-reference (SR), SR-CC theory is notoriously hard to spin-adapt for such cases,³¹ not to mention CC's MR generalizations. Satisfying this spin property automatically is a major computational and conceptual simplification.

In this paper, we present the explicit equations to calculate DA-EOM-CC and MR-DA-CC. Unlike other MR-CC methods, the method lends itself to “black-box” application. We illustrate the methods for several examples of bond-breaking into radical products, while exclusively using restricted Hartree–Fock (RHF) references and proper spin-eigenfunctions. We also address the question of excitation energies for open-shell atoms and the notorious example of twisted ethylene. Because of the favorable treatment of essential effects of triple excitations, MR-DA-CCSD \approx MR-DA-CCSDT provides a very economical description of traditional multireference problems with scaling no worse than CCSD itself.

THEORY

In the coupled cluster theory, the wave function $|\Psi_0\rangle$ is defined via the exponential Ansatz

$$|\Psi_0\rangle = e^T |\Phi_0\rangle \quad (1)$$

Received: March 22, 2011

Published: September 14, 2011

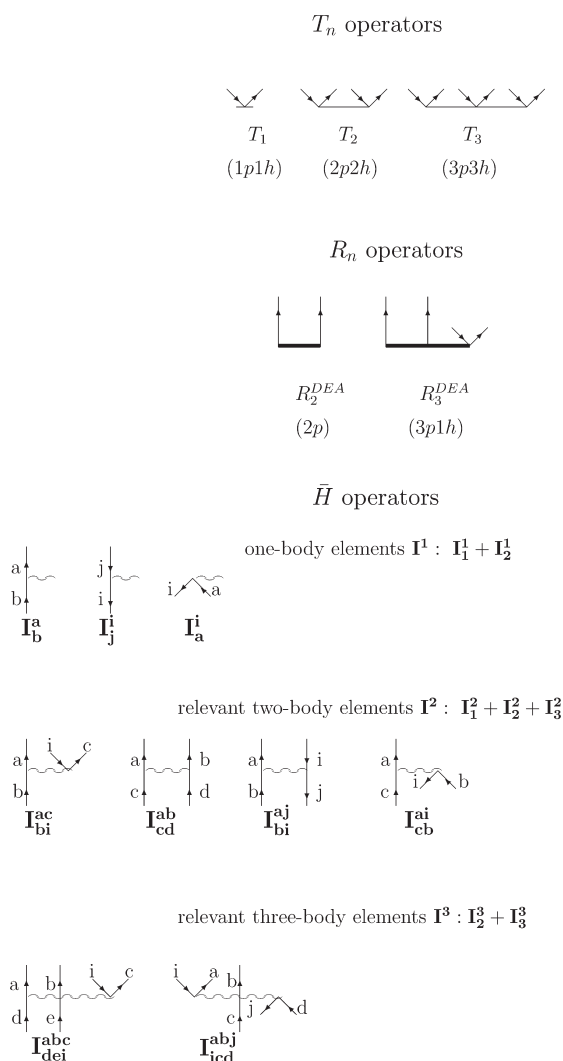


Figure 1. Diagrammatic form of the T , R , and \bar{H} (i.e., I_k^n elements: n -body with k annihilation lines) operators used in the MR-DA-CCSDT method. Types of lines used in the definition of the T and R operators: p , particle; h , hole.

where T is a cluster operator and $|\Phi_0\rangle$ is the reference determinant. Within the CCSDT model, the cluster operator is limited to triple excitations $T = T_1 + T_2 + T_3$ with the usual definition: $T_k = (k!)^{-2} \sum_{ab\dots} \sum_{ij\dots} t_{ij\dots}^{ab\dots} \{a^\dagger b^\dagger \dots j i\}$, where the indices $a, b, \dots (i, j, \dots)$ refer to virtual (occupied) one-particle levels. The Schrödinger equation is written in the form $H_N |\Psi_0\rangle = \Delta E_0 |\Psi_0\rangle$, where $H_N = H - \langle \Phi_0 | H | \Phi_0 \rangle$ and ΔE_0 is the correlation energy. The cluster amplitudes, $t_{ij\dots}^{ab\dots}$, are obtained by solving the CC equations, $\langle \Phi_{ij\dots}^{ab\dots} | (H_N e^T)_c | \Phi_0 \rangle = 0$.

Within the EOM formalism, the k -state wave function ($|\Psi_k\rangle$) is obtained by the action of an $R(k)$ operator on the ground state (GS) wave function ($|\Psi_0\rangle$):

$$|\Psi_k\rangle = R(k) |\Psi_0\rangle \quad k = 1, 2, \dots \quad (2)$$

The $R(k)$ is linear (CI-like), and its choice depends on the particular process. For the double electron attachment, it can be expressed as (see Figure 1 for the diagrammatic form of it):

$$R(k) = R_2(k) + R_3(k) \quad (3)$$

where we limit the cluster expansion to the double and triple components (the R_1 part cannot be defined within the DA formalism). The $R(k)$ operator can be expressed through the elementary creation–annihilation operators as

$$R(k) = \frac{1}{2} \sum_{ab} r^{ab}(k) \{a^\dagger b^\dagger\} + \frac{1}{6} \sum_{abc} \sum_i r_i^{abc}(k) \{a^\dagger b^\dagger c^\dagger i\} \quad (4)$$

Thus, for the double electron attachment problem, we create electrons in levels a and b ($R_2(k)$), or in addition, we add single excitation from i to c ($R_3(k)$). These single excitations are sometimes called S_1 in the following. The effect of $\{a^\dagger b^\dagger\}$ working on an $n - 2$ electron vacuum, which is a kind of “core” vacuum, generates four determinants composed of two “valence” orbitals that can become quasi-degenerate, thereby naturally introducing the four determinant MR description of the problem. Then, S_1 accounts for single excitations among these four determinants, analogous to adding singles to a MR-CI. Such terms are extremely important. In principle, further excitations could be made from the “CI” side of the Ansatz. The choice of orbitals to use in the $n - 2$ electron vacuum is arbitrary, and many different choices were considered in our prior paper.¹⁹ But, the obvious choice for the DA case is HF core orbitals, since these do not have most of the pathologies that can be encountered in the DI problem.

The Schrödinger equation for the double electron-attached states ($k = 1, 2, \dots$) can be written as

$$H_N R(k) |\Psi_0\rangle = \Delta E_k R(k) |\Psi_0\rangle \quad (5)$$

Taking advantage of the definition of the CC wave function, eq 1, after simple algebra we arrive at the final EOM equation in the form

$$\bar{H}_N R(k) |\Phi_0\rangle = \omega_k R(k) |\Phi_0\rangle \quad (6)$$

where \bar{H}_N is the similarity transformed Hamiltonian defined as

$$\bar{H}_N = e^{-T} H e^T - \langle \Phi_0 | e^{-T} H e^T | \Phi_0 \rangle \quad (7)$$

and the $\omega_k = \Delta E_k - \Delta E_0$ is the energy change connected with the double electron-attachment process.

In the matrix form, eq 6 can be written as

$$\bar{H}_N \mathbf{R}(\mathbf{k}) = \omega_k \mathbf{R}(\mathbf{k})$$

A diagonalization of the matrix \bar{H}_N in the subspace of double-electron attached configurations provides required energies and eigenvectors.

As is known,¹² the single electron-attachment R operator (presented as linear in the EOM formalism) can be obtained also in the linear form via exponential expansion introduced in the Fock space CC (FS-CC) theory,^{32–39} since all products of $S^{(1,0)}$ in $\exp(S^{(1,0)})$ vanish, leaving just $S^{(1,0)} = R$. This makes EA-EOM-CC entirely linked, connected, extensive, and intensive for principal EAs. For EE-EOM-CC, all excited states are intensive, but since $R(k)$ does not correspond to $\exp(S^{(1,1)})$,^{32,34,35,40} the equations are not composed solely of linked diagrams. The same refers to the double attached states (two valence sector in FS formalism: (2,0)). But if deemed important, as shown in ref 41, there is a possibility to eliminate this drawback of the EE-EOM-CC theory via EOM-CCx (size-extensive EOM-CC) strategy based on the Canonical Bloch Equation of FS-CC,⁴² but at the cost of introducing a choice of “active orbitals”. We do not want to do that here, since the transparency of the MR-DA-CC theory to active orbitals is one of its primary advantages.

It is well-known that $\bar{\mathbf{H}}_N$ is a non-Hermitian matrix, so there is a left-hand eigenvector, $\mathbf{L}(\mathbf{k})\bar{\mathbf{H}}_N = \omega_{\mathbf{k}}\mathbf{L}(\mathbf{k})$. Both $\mathbf{L}(\mathbf{k})$ and $\mathbf{R}(\mathbf{k})$ eigenvectors have the same $\omega_{\mathbf{k}}$ eigenvalue, and both are needed to obtain ordinary and transition density matrices.

Since the EOM scheme can be viewed in a CI-like way, it can be reduced to the diagonalization of a matrix within the appropriate configurational subspace. At the CCSD level for the DA case, we diagonalize a smaller matrix than for the EA problem:¹²

$$\bar{\mathbf{H}} = \langle \mathbf{D} | \bar{\mathbf{H}} | \mathbf{D} \rangle$$

where $D \equiv \Phi^{ab}$ represents the Slater determinants corresponding to the double electron attached configurations. At the “triples” level, we would have an additional configuration of the type $T \equiv \Phi_i^{abc}$, i.e., double attachment is accompanied by single excitation, S_1 , contributing to R :

$$\bar{\mathbf{H}} = \begin{bmatrix} \langle \mathbf{D} | \bar{\mathbf{H}} | \mathbf{D} \rangle & \langle \mathbf{D} | \bar{\mathbf{H}} | \mathbf{T} \rangle \\ \langle \mathbf{T} | \bar{\mathbf{H}} | \mathbf{D} \rangle & \langle \mathbf{T} | \bar{\mathbf{H}} | \mathbf{T} \rangle \end{bmatrix}$$

When we use a CCSD reference state but add R_3 terms into R , we get most of the benefits of triples without ever having to build the CCSDT solution, which would add an $\sim n_{\text{occ}}^3 n_{\text{vir}}^5$ step in its evaluation. That mixture of terms leads to the MR-DA-CCSD viewpoint.

MR-DA-CC Working Equations. In this section, we present explicit working equations for the MR-DA-CC method. This requires solution of the standard CC equations to provide the T_n amplitudes and then the construction of the \bar{H} operator, according to eq 7. To obtain the eigenvalues and eigenvectors of the \bar{H} operator, we employ a direct diagonalization scheme for non-Hermitian matrices⁴³ analogous to the standard Davidson method.⁴⁴ When the right-hand solutions are sought, the Davidson scheme requires the product of the MR-DA-CC matrix right multiplied by the amplitude vector R , i.e., $\bar{H}R$. When R and T correspond to the same level of excitation, MR-DA-CC = DA-EOM-CC

We consider two models for the DA problem based on the CCSD and CCSDT solutions. The former requires determination of the R_2 (i.e., $2p$; p , particle) amplitudes only and the latter, both R_2 (i.e., $2p$) and R_3 (i.e., $3p1h$; h , hole).

DA-EOM-CC is exactly MR-DA-CCSD without S_1 and is quite simple. It is presented in Figure 2a in a diagrammatic form and in Table 1 (in the CCSD section) in an algebraic form. At the CCSDT level, the DA equations are much more complicated, see Figure 2b or the lower part of Table 1. Note that to each term in Table 1 the proper permutation of the external indices should be applied, as indicated by the symbol $P(\dots/\dots)$ (see the explanation given in the footnote to Table 1). It should be explained also that in Tables 1 and 2 we adopt the tensor notation, which implies summation over repeated indices.

The form of \bar{H} required for the construction and solution of the MR-DA-CCSDT equations in their standard form involves up to three-body elements. Emphasizing the many-body structure of \bar{H} , we may decompose it into individual contributions as follows:

$$\bar{H} = I^1 + I^2 + I^3 \quad (8)$$

We may rewrite the particular element, I^n , as the sum over components I_k^n , where k indicates the number of annihilation

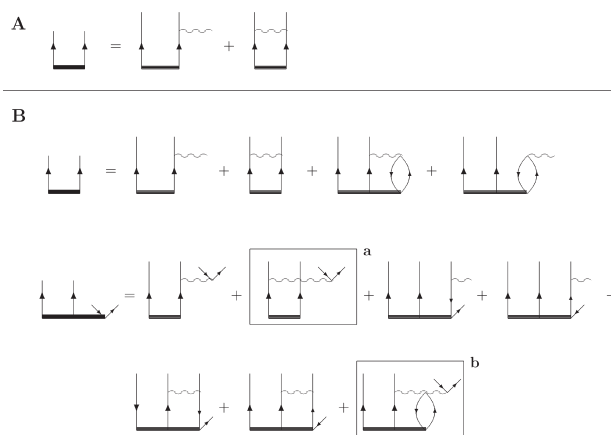


Figure 2. Diagrammatic form of the MR-DA-CC equations (A, MR-DA-CCSD no S_1 variant; B, MR-DA-CCSDT variant) in antisymmetrized formalism for the standard version.

Table 1. MR-DA-CCSD no S_1 and MR-DA-CCSDT Equations in Goldstone Formalism for Standard and Factorized Version

expression ^a
MR-DA-CCSD no S_1
$(\bar{H}R)^{ab} = P(a/b)[r^{ae}t_e^b + 1/2r^{ef}t_{ef}^{ab}]$
MR-DA-CCSDT
$(\bar{H}R)^{ab} = P(a/b)[r^{ae}t_e^b + 1/2r^{ef}t_{ef}^{ab} + 2r_m^{aef}t_{ef}^{bmn} - r_m^{aef}t_{ef}^{mb} - r_m^{efb}t_{ef}^{am} + r_m^{abc}t_m^{bc} - r_m^{ab}t_m^{bc}]$
$(\bar{H}R)^{abc} = P(a/b)[r^{ae}t_e^{bc} + 1/2r_i^{abc}t_{ci}^{bc} + 1/2r_m^{abc}t_m^{bc} + r_i^{abc}t_{ci}^{bc} + 1/2r_i^{efc}t_{ef}^{abc} + r_i^{aef}t_{ef}^{abc} - r_m^{acc}t_m^{bcm} - 1/2r_m^{abc}t_m^{bc} + r_m^{abc}t_m^{bc} - r_m^{abc}t_m^{bc}] + F_i^{abc}$
$F_i^{abc} = P(a/b)[1/2r_m^{aef}t_{ef}^{abc} + 2r_m^{aef}t_{ef}^{bmc} - r_m^{aef}t_{ef}^{mbc} - r_m^{efb}t_{ef}^{amc}]$
$F_i^{abc} = P(a/b)[1/2r_m^{abc}t_m^{bc} - r_m^{abc}t_m^{bc}]$

Table 2. Algebraic Expression for the χ Intermediates Used in the MR-DA-CCSDT Model

intermediate	expression ^a
χ^{ij}	$+ r^{ef}v_{ef}^{ij}$
χ^{ai}	$+ r^{ef}t_{ef}^{ai} + 2r_m^{aef}v_{ef}^{im} - r_m^{efa}v_{ef}^{im} - r_m^{aef}v_{ef}^{mi}$

^a Summation over repeated indices assumed; summation indices belong to the set $\{e, f, m\}$.

lines (see Figure 1):

$$\begin{aligned} I^1 &= I_1^1 + I_2^1 \\ I^2 &= I_1^2 + I_2^2 + I_3^2 \\ I^3 &= I_2^3 + I_3^3 \end{aligned} \quad (9)$$

We skipped the \bar{H} components with 0 annihilation lines since they cannot be contracted with the R operators. Similarly, in the last equation, the I_1^3 term is omitted since it does not enter the MR-DA-CCSDT model. The same refers to the I_4^2 element of \bar{H}

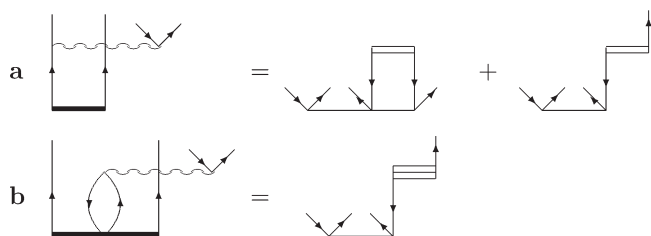


Figure 3. Diagrammatic notation of the factorization of selected contributions to the MR-DA-CC equations.

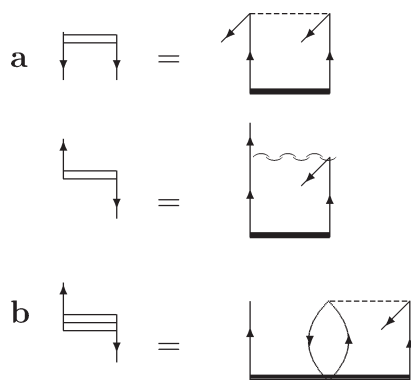


Figure 4. Diagrammatic form of the intermediates used in the MR-DA-CCSDT equations in antisymmetrized formalism.

in the second equation. We apply here our diagrammatic formalism, which is discussed, for example, in refs 13 and 14.

The many-body type of the \bar{H} elements expressed through the $I_{r,\dots}^s$ amplitudes are determined by the number of indices, i.e., I_r^s , I_{tw}^s and I_{uvw}^s corresponding to the I^1 , I^2 , and I^3 operators, respectively. Note that the r,s,\dots symbols indicate general indices being of either hole or particle character.

It is easy to also assign creation–annihilation character to the particular $I_{s,\dots}^r$ amplitude. The index referring to the annihilation (creation) line is represented by the presence of the hole (particle) symbol as a superscript or the particle (hole) symbol as the subscript. To clarify this point, we give the following definition of the I_k^r elements expressed through their regular, antisymmetrized form:

$$\begin{aligned}
 I_1^r &= \sum_{a,b} I_{ab}^r a^+ b + \sum_{i,j} I_{ij}^r i^+ j \\
 I_2^r &= \sum_{a,i} I_a^r i^+ a \\
 I_1^2 &= \sum_{a,b,c,i} I_{bi}^{ac} a^+ c^+ i b \\
 I_2^2 &= \sum_{a,b,c,d} I_{cd}^{ab} a^+ b^+ d c + \sum_{a,b,i,j} I_{bi}^{aj} a^+ j^+ i b \\
 I_3^2 &= \sum_{a,b,c,i} I_{cb}^{ai} a^+ i^+ b c \\
 I_2^3 &= \sum_{a,b,c,d,e,i} I_{dei}^{abc} a^+ b^+ c^+ i e d \\
 I_3^3 &= \sum_{a,b,c,d,i,j} I_{icd}^{abj} a^+ b^+ j^+ d c i
 \end{aligned} \quad (10)$$

All contributions to the elements of the \bar{H} operator used in the MR-DA-CC models proposed in this work are defined in refs 11 and 12.

The MR-DA-CC equations assume that we use all required \bar{H} elements regardless of the complexity of a particular term. This means that we employ all appropriate three-body terms. That form of the MR-DA-CCSDT equations is presented diagrammatically in Figure 2b and—in its algebraic form—in the b equations in Table 1. Such a formulation of the MR-DA-CC problem, although the most natural one, would result in higher scaling of the computational procedure either in the \bar{H} construction or in the solution of the EOM-CC equations. To avoid this, we apply a factorization scheme which naturally accounts for the difficult terms that involve the three-body \bar{H} elements. These terms are indicated in Figure 2b by the rectangles, and their factorized construction is explained in Figures 3 and 4. Hence, the diagram indicated by the letter a in the R_3 equation in Figure 2b is replaced by the contribution given in Figure 3a, where the heavy horizontal line corresponds to the R amplitude and the thin one to the T amplitude (see also Figure 1 for definition of the operators). Thus, we avoid using any explicit three-body \bar{H} element in the R_3 equation by construction of the intermediate shown in Figure 4a and then by employing it in the diagram on the right-hand side of the equation in Figure 3a. The same procedure is applied to diagram b in Figure 2b. By the intermediate, we understand throughout the paper the quantity obtained by the contraction of the integral or \bar{H} element with the R operator.

The factorization procedure shown diagrammatically in Figures 3 and 4 is shown algebraically in Table 1, where the contribution denoted by F_i^{abc} represents the terms corresponding to the diagrams indicated by rectangles in Figure 2b. The expression for F_i^{abc} is constructed in a standard, nonfactorized way, engaging only \bar{H} elements and R amplitudes, see $F_i^{abc b}$, or in a factorized way, see $F_i^{abc c}$. The latter requires construction of the intermediates χ defined in Table 2.

The factorization procedure makes the evaluation of the ($\bar{H}R$) quantities much more efficient. We should remember, however, that the \bar{H} elements are computed only once in the whole process, while the intermediates must be recomputed in each iteration. However, that is a most efficient step.

Approximations. As we can see, the equations for the DA amplitudes are much simpler than those for the EA case.¹² In addition, the DA part scales only as $n^6(n_{\text{occ}}^1 n_{\text{vir}}^5)$ at the CCSDT level. We also do calculations using the MR-DA-CCSD variant in which the GS is the CCSD one (scaling $n^6(n_{\text{occ}}^2 n_{\text{vir}}^4)$) instead of CCSDT (scaling $n^8(n_{\text{occ}}^3 n_{\text{vir}}^5)$), and in the EOM part we take R_2 and R_3 as in the full CCSDT model. For comparison purposes, we also implement DA-EOM-CCSD (=MR-DA-CCSD no S_1 variant). The DA part scales as $n^4(n_{\text{vir}}^4)$, as it contains only the R_2 equation (first two diagrams in Figure 2). Finally, we consider an additional variant in which we replace the GS CCSDT^{45,46} with an approximate version CCSDT-3,⁴⁷ termed, MR-DA-CCSDT-3. This variant relies on employing a rigorous treatment of R_2 and R_3 with the ground state CC scaling as $n_{\text{occ}}^3 n_{\text{vir}}^4$. The approximate variant behaves in the same manner as its rigorous counterpart (see the Results section), and it can be applied for larger systems for which the effect of triples is essential. We may use the approximate MR-DA-CCSDT-3=DA-EOM-CCSDT-3 method without significantly lowering the accuracy.

RESULTS

A. Twisted Ethylene. The newly implemented methods are applied for the widely known twisted ethylene case⁴⁸

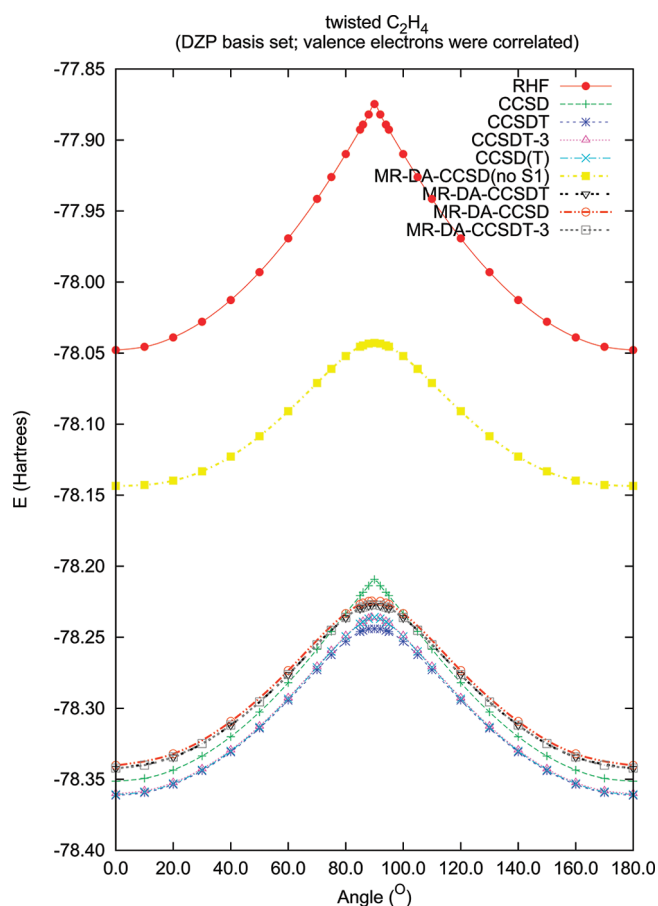


Figure 5. The RHF, CC, and MR-DA-CC torsional curves for the twisted ethylene in DZP basis set and valence electrons correlated ($R_{CC} = 1.3390 \text{ \AA}$, $R_{CH} = 1.0856 \text{ \AA}$, $\angle_{HCH} = 117.6^\circ$).

(see Figures 5 and 6), which has a multireference effect. In the calculations, we use DZP⁴⁹ and cc-pVTZ⁵⁰ basis sets with valence electrons correlated and the experimental geometry $R_{CC} = 1.3390 \text{ \AA}$, $R_{HC} = 1.0856 \text{ \AA}$, and $\angle_{HCH} = 117.6^\circ$. The energy is plotted as a function of the dihedral angle between the two methylene (CH_2) groups. At the very top, we have the SCF curves which show the cusp at 90° . The same is shown by the standard CCSD model, but when using the DA strategy we obtain a smooth curve in the critical region of the potential energy curve (PEC). The same effect is observed for both basis sets employed; the cc-pVTZ curves are slightly shifted toward lower energy, but the two sets are parallel. When adding triples, the DA curves are of the same shape and without the typical cusp retained in CCSD total energy. The DA calculations are performed in such a way that we do ground state CC calculations for the double cation, and the neutral molecule energy is obtained with DA eigenvalues. For the planar conformation, we have no problem with applying any quantum chemical method, but when the dihedral angle is changed to the 90° conformation, then the HOMO–LUMO gap becomes small, and we encounter various divergencies, demonstrating its MR character. To the contrary, when we do calculations for the double-electron attached molecule (alternatively for the double ionized molecule¹⁹), the problem disappears. Thus, we can use MR-DA-CC methods to obtain an accurate description of some of the traditional multireference situations.

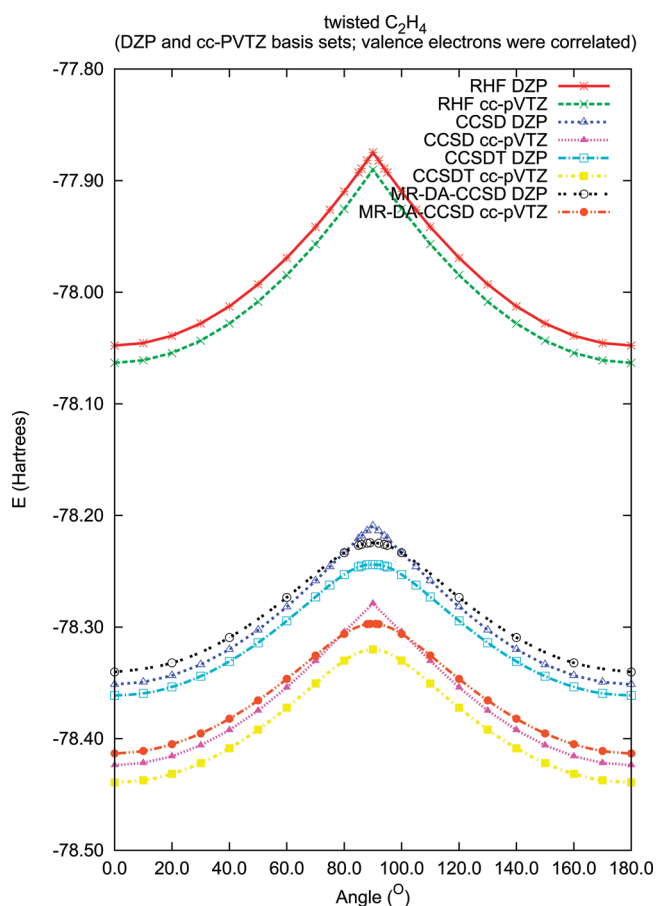


Figure 6. The RHF, CCSD, CCSDT, and MR-DA-CCSD torsional curves for the twisted ethylene in DZP and cc-pVTZ basis sets and valence electrons correlated ($R_{CC} = 1.3390 \text{ \AA}$, $R_{CH} = 1.0856 \text{ \AA}$, $\angle_{HCH} = 117.6^\circ$).

B. Sodium Dimer. The next example of the application of the new method is the PEC for the Na_2 molecule (see Figures 7 and 8). The calculations use the POLI⁵¹ basis set, and the $1s$ orbitals are kept frozen. In the $n - 2$ reference, the Na_2 system is a closed shell one. Doing calculations around the equilibrium using standard methods is straightforward, but when we try to do, for example, the whole potential energy curve, then for larger R we may have problems with obtaining the reference function since Na_2 would dissociate into two open-shell fragments (Na atoms). However, when we do calculations using the DA strategy, we do it for the Na_2^{2+} cation, which dissociates into two closed shell fragments (isoelectronic with neon atom). The essence of this trick relies on the fact that DA calculations for the Na_2^{2+} for large R are much easier to carry to complete separation as spin-eigenstates. In Figure 7, we see that all curves representing the MR-DA methods based on the full R_3 approximation stay close to the CCSDT curve (which is considered here a reference one) irrespective of the CC model used for the GS. Incorrect behavior is observed for the R_2 only model (no R_3 ; top curve) and also for the noniterative CCSD(T) curve, showing a hump around 6 \AA .

In Figure 8, we show the nonparallelity error curve representing the differences between the reference CCSDT values and the MR-DA-CCSD ones. Nearly within the whole range of the internuclear distances, they are very stable and small (ca. 3.6 mH),

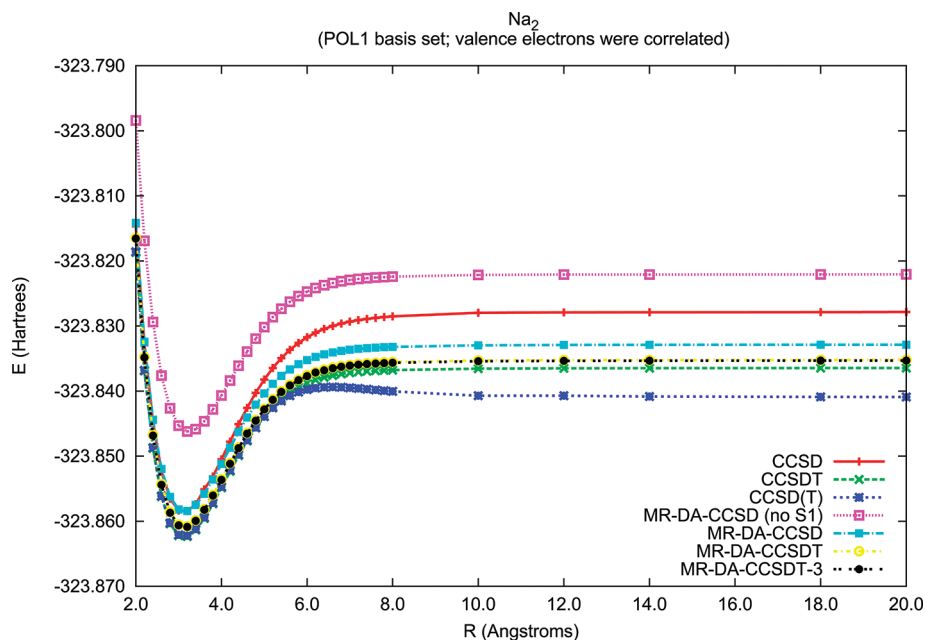


Figure 7. Potential energy curves for the Na_2 molecule with MR-DA-CC and CC methods.

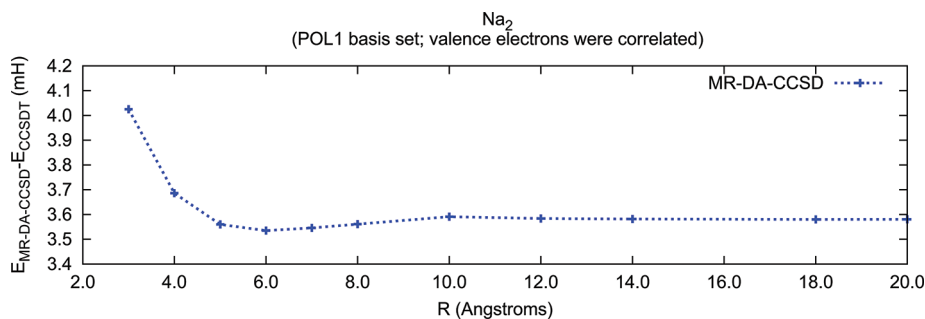


Figure 8. Energy difference from CCSDT for Na_2 ground state MR-DA-CCSD.

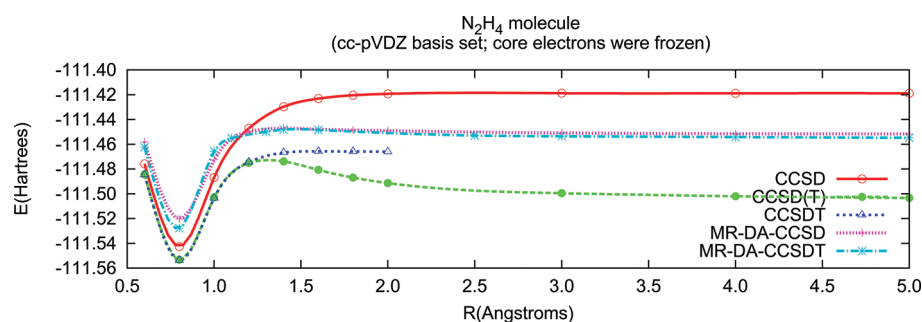


Figure 9. The CC and MR-DA-CC curves for the N_2H_4 molecule in the cc-pVDZ basis set and valence electrons correlated ($R_{\text{NH}} = 1.016 \text{ \AA}$, $\angle_{\text{HNN}} = 108.85^\circ$, $\angle_{\text{HNN}} = 106^\circ$).

going up to 4.0 mH at short R values. As will be presented elsewhere, this underbinding near equilibrium does not happen with the MR-DI-CCSD version. We attribute the differing results to defects of using the double cation orbitals in the MR-DA-CCSD results presented here, but there are also other aspects to consider. In the bond-breaking MR regions of the curves, though, MR-DA-CCSD is excellent and seems to be better than its MR-DI-CCSD counterpart.

In the context of the Na_2 dissociation, we should mention the size-extensivity issue. A characteristic feature of the EOM theory is its size-intensivity, which states that upon separation of two fragments ($AB^* \rightarrow A^* + B$) the energies corresponding to the local excitation (or local double ionization or local double-electron attachment) are size-extensive. When the process in question engages both fragments, e.g., a charge-transfer excitation in EE-EOM-CC, then the equations do not correspond to a

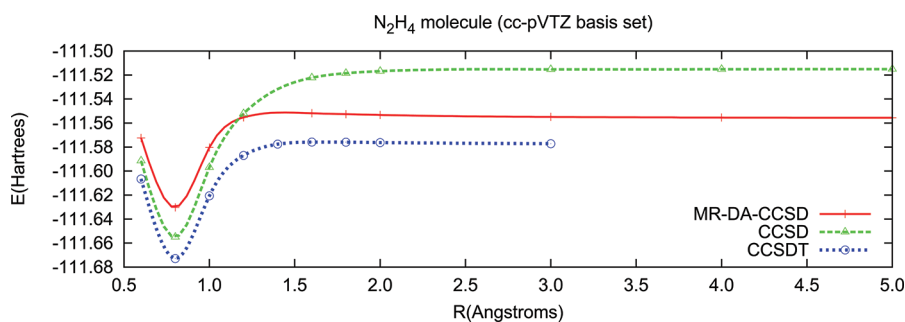


Figure 10. The CCSD and MR-DA-CCSD curves for the N_2H_4 molecule in the cc-pVTZ basis set (frozen orbitals: first two occupied and 11 highest unoccupied ones; $R_{NH} = 1.016 \text{ \AA}$, $\angle_{HNN} = 108.85^\circ$, $\angle_{HNH} = 106^\circ$).

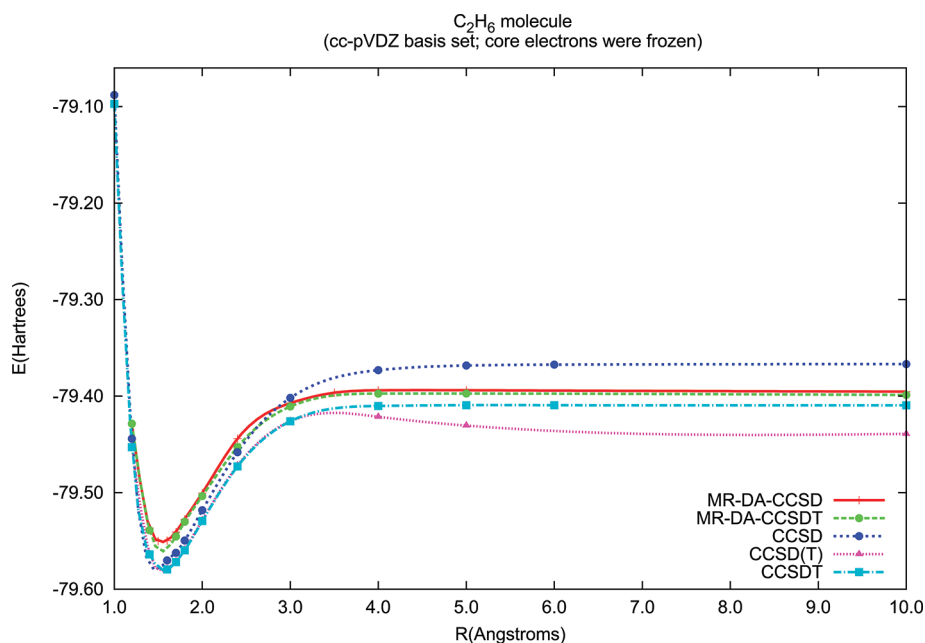


Figure 11. The CC and MR-DA-CC curves for the C_2H_6 molecule in the cc-pVDZ basis set and valence electrons correlated ($R_{CH} = 1.091 \text{ \AA}$, $\angle_{HCC} = 110.91^\circ$, $\angle_{HCH} = 108^\circ$).

fully linked form, and the dissociation of the Na_2 ground state belongs to this category. This feature occurs because the terms involved for both fragments correspond to higher excitations than are included in the level of the calculation like CCSD. So this is not as much a failure of the theory as its truncation to singles and doubles, for example. The question of size-intensive quantities being the difference between two size-extensive ones, or just size-intensive without that condition, is an interesting one.

C. Hydrazine. In Figure 9, we show PECs for the N–N bond of the N_2H_4 molecule with three computational methods of SR character (CCSD, CCSDT, and CCSD(T)) and two MR-DA based on the CCSD and CCSDT reference obtained for the cc-pVDZ⁵⁰ basis set. In Figure 10, we plot three potential energy curves obtained for CCSD, CCSDT, and MR-DA-CCSD schemes using a cc-pVTZ basis set.⁵⁰ For the CCSDT in both basis sets, the convergence can be reached only for R values no longer than 2–3 Å. For the CCSD scheme, the CC equations converge within the whole range of R values, but the curve is too high. The CCSD(T) shows its typical failure with a hump around 1.2 Å. The MR-DA curves show correct behavior for both basis sets and for two types of GS references. The obvious inadequacy

of the MR curves is a too high energy around equilibrium that we believe can be remedied, partly by employing orbitals optimized for the n electron problem instead of the $n - 2$ one used in the current work.

D. Ethane. Similar observations can be made with respect to the dissociation of the C–C bond and resulting PECs for the C_2H_6 molecule, see Figures 11 and 12. We show the same set of PECs as for the N_2H_4 molecule both for the cc-pVDZ (Figure 11) and cc-pVTZ (Figure 12) basis sets. The observations are similar: both MR curves shown in Figure 11 stay close to the CCSDT reference, and the same is true for the MR-DA-CCSD curve shown in Figure 12. In the case of the C_2H_6 system, the single reference CCSDT curve can be obtained for the whole range of R values. Both the SR CCSD curves (Figures 11 and 12) deviate toward higher energies, which visibly worsens the performance of the CCSD model. Though the results for N_2H_4 and C_2H_6 molecules are qualitatively similar at all MR levels, the MR-DA curves based on the CCSD reference are computed with significantly lower computer effort than those based upon CCSDT.

MR-DA results can be seen as an alternative route to evaluate potential energy curves for the ground and excited states. The

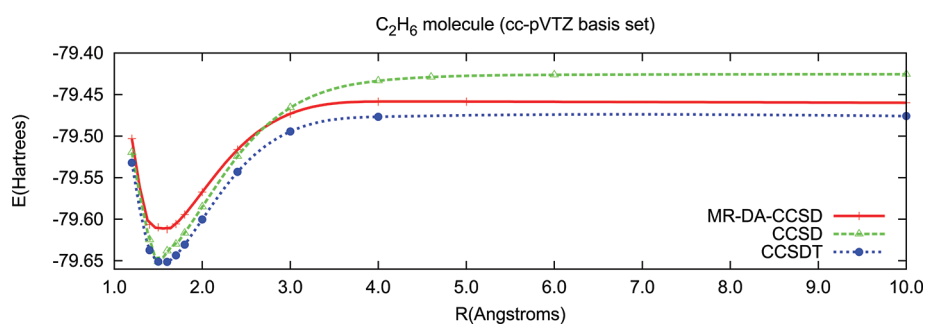


Figure 12. The CCSD and MR-DA-CCSD curves for the C_2H_6 molecule in the cc-pVTZ basis set (frozen orbitals: first two occupied and 41 highest unoccupied ones; $R_{CH} = 1.091 \text{ \AA}$, $\angle_{HCC} = 110.91^\circ$, $\angle_{HCH} = 108^\circ$).

Table 3. Excitation Energies (eV) for the C Atom Using the MR-DA-CC Methods in the POL1 Basis Set

sym.	MR-DA-CC					exptl ^b
	SD (no S_1)	SDT	SD	SDT-3	AEL ^a	
	$T = T_1 + T_2^c$ $R = R_2$	$T = T_1 + T_2 + T_3^d$ $R = R_2 + R_3$	$T = T_1 + T_2^c$ $R = R_2 + R_3$	$T = T_1 + T_2 + T_3^e$ $R = R_2 + R_3$		
$2p^2 \ ^3P$	0.0	0.0	0.0	0.0	91.5%	0.0
$2p^2 \ ^1D$	1.308	1.480	1.481	1.480	92.7%	1.26
$2p^2 \ ^1S$	1.771	2.802	2.805	2.801	94.5%	2.68
$2p3s \ ^3P^0$	5.834	7.318	7.320	7.318	95.4%	7.48
$2p3s \ ^1P^0$	5.963	7.520	7.521	7.520	96.7%	7.68
$2p3p \ ^1P$	7.107	8.503	8.504	8.503	95.9%	8.54
$2p3p \ ^3D$	7.228	8.670	8.671	8.670	96.1%	8.64
$2p3p \ ^3S$	7.411	8.869	8.870	8.869	96.1%	8.77
$2p3p \ ^3P$	7.748	9.026	9.027	9.026	94.9%	8.85
$2p3p \ ^1D$	8.051	9.382	9.384	9.382	96.1%	9.00
MAE ^f	1.175	0.154	0.155	0.154		

^a Approximate excitation level indicates % of the R_2 (i.e., $2p$) configurations. ^b ref 53. ^c GS: CCSD. ^d GS: CCSDT. ^e GS: CCSDT-3. ^f MAE: mean absolute error.

method clearly provides superior bond-breaking subject to RHF reference functions. The results tend to fall between SR-CCSD and SR-CCSDT, but with none of the expense of the latter and better bond breaking behavior with RHF reference functions in either case.

E. Carbon Excitations. Finally, we apply DA calculations to determine the excitation spectra. Though it is planned to treat a number of transition metal atoms in this way, we first illustrate it for the carbon atom which has two electrons out of a closed shell structure. Adding two electrons to the Be configuration of C^{2+} , we can obtain excited states of carbon (see Table 3) in a very economical way. In the calculations, we use the POL1 basis set with 6 d functions and all electrons correlated. We quote in Table 3 the excitation energies for nine low lying states, relating them to the experimental data. Four different methods are considered: one (CCSD) based on the CCSD solution for the ground state and on the EOM-CCSD equation for the DA part. The other methods engage R_3 in the results. The most important column is the third, where the reference state is the CCSD ($n_{occ}^2 n_{vir}^4$) solution and the target states solutions at the DA level are ($\sim n_{occ}^1 n_{vir}^5$), which means single excitations of the MR space generated by the DA operator, $\{a^\dagger b^\dagger\}$. Its inclusion provides nearly identical results to the full inclusion of triples, a remarkable improvement over the DA-EOM-CCSD (=MR-DA-CCSD no S_1)

value equal to 1.175 eV (see second column in Table 3). Clearly, the quality of the ground state wave function is less important than that for the target state. The results also compare well with experimental results. In most cases, the deviation remains between 0.04 and 0.20 eV. It grows, however, for the higher energy states.

In Table 3, we also list the AEL (approximate excitation level) values for the computed states of the C atom. In the last column are the percentages of the R_2 type configurations in the computed double-electron attached states. All of the states obtained are dominated by the configurations created by the R_2 operator.

CONCLUSIONS

A new MR-DA-CC method is introduced and applied with full inclusion of the connected triple excitations. When R and T include the same excitation levels, this is simply DA-EOM-CC, but by mixing them with an eye toward MR excitations in the target state, much better results are obtained, and at substantially less cost than the unmodified DA-EOM-CCSDT. By adopting the proper factorization strategy, all terms involving the evaluation of the three-body elements of the \bar{H} operators are bypassed by a rigorous factorization procedure. The factorization makes it possible to achieve a scaling of the target state that is no worse

than $n_{\text{occ}}^1 n_{\text{vir}}^5$ after the ground state is obtained. All of the resulting equations are presented diagrammatically and algebraically with a detailed presentation of the factorization scheme.

The program developed, part of ACES II,⁵² has been subsequently applied to the calculation of the DA for production of the PEC for some systems when double cation separates into two closed shell fragments. The bottleneck of the calculation occurs in the GS calculations. To address this, we propose some approximations with better scaling that yield comparable results to these for the full or more complete methods.

The MR-DA-CC methods are highly suitable for doing calculations for systems that require a traditional multireference description in a very economical, unbiased, and routine way. The double attachment scheme can be used to calculate excitation spectra of open-shell systems. It can be applied to describe systems that have two electrons out of a closed shell structure (i.e., difficult transition states, biradicals, etc.), and it will be generalized to three, four, and more attached states, TA, QA, etc. The approach potentially offers a very attractive, easily applied MR-CC method for many problems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: musial@ich.us.edu.pl

ACKNOWLEDGMENT

This work has been supported by the Ministry of Science and Higher Education, Poland, under grant No. N N204 090938, and also by the United States Air Force Office of Scientific Research under STTR AF09-BT40.

REFERENCES

- (1) Rowe, D. J. *Rev. Mod. Phys.* **1968**, *40*, 153.
- (2) Sekino, H.; Bartlett, R. J. *Int. J. Quantum. Chem. Symp.* **1984**, *18*, 255–265.
- (3) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57–62.
- (4) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029–7039.
- (5) Comeau, D. C.; Bartlett, R. J. *Chem. Phys. Lett.* **1993**, *207*, 414–423.
- (6) Emrich, K. *Nucl. Phys. A* **1981**, *351*, 379–396.
- (7) Kucharski, S. A.; Wloch, M.; Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2001**, *115*, 8263–8266.
- (8) Kowalski, K.; Piecuch, P. J. *Chem. Phys.* **2001**, *115*, 643–651.
- (9) Korona, T.; Werner, H. J. *J. Chem. Phys.* **2003**, *118*, 3006–3019.
- (10) Hirata, S. *J. Chem. Phys.* **2004**, *121*, 51–59.
- (11) Musial, M.; Kucharski, S. A.; Bartlett, R. J. *J. Chem. Phys.* **2003**, *118*, 1128–1136.
- (12) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2003**, *119*, 1901–1908.
- (13) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352 and references therein.
- (14) Shavitt, I.; Bartlett, R. J. *Many-Body Methods in Chemistry and Physics: Many-Body Perturbation Theory and Coupled Cluster Methods*; Cambridge Press: Cambridge, England, 2009.
- (15) Krylov, A. I. *Annu. Rev. Phys. Chem.* **2008**, *59*, 433–462.
- (16) Musial, M.; Kucharski, S. A.; Bartlett, R. J. *Adv. Quantum Chem.* **2004**, *47*, 209–222.
- (17) Musial, M. *Mol. Phys.* **2010**, *108*, 2921–2931.
- (18) Hirata, S.; Nooijen, M.; Bartlett, R. J. *Chem. Phys. Lett.* **2000**, *326*, 255–262.
- (19) Musial, M.; Perera, A.; Bartlett, R. J. *J. Chem. Phys.* **2011**, *134*, 114108–1–10.
- (20) Sattelmeyer, K. W.; Schaefer, H. F., III; Stanton, J. F. *Chem. Phys. Lett.* **2003**, *378*, 42–46.
- (21) Demel, O.; Shamasundar, K. R.; Kong, L.; Nooijen, M. *J. Phys. Chem. A* **2008**, *112*, 11895–11902.
- (22) Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1997**, *107*, 6812–6830.
- (23) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2011**, *135*, 044121–1–8.
- (24) Chaudhuri, R.; Datta, B.; Das, K.; Mukherjee, D. *Int. J. Quantum Chem.* **1996**, *60*, 347–358.
- (25) Ida, T.; Ortiz, J. V. *J. Chem. Phys.* **2008**, *129*, 084105–1–11.
- (26) Schulte, H. D.; Cederbaum, L. S. *J. Chem. Phys.* **1996**, *105*, 11108–11133.
- (27) Van Huis, T. J.; Wesolowski, S. S.; Yamaguchi, Y.; Schaefer, H. F., III. *J. Chem. Phys.* **1999**, *110*, 11856–11864.
- (28) Tobita, M.; Perera, S. A.; Musial, M.; Bartlett, R. J.; Nooijen, M.; Lee, J. S. *J. Chem. Phys.* **2003**, *119*, 10713–10723.
- (29) Eliav, E.; Kaldor, U.; Ischikawa, Y. *Phys. Rev. A* **1996**, *53*, 3050–3056.
- (30) Nayak, M. K.; Chaudhuri, R. K. *Eur. Phys. J.* **2006**, *37*, 171–176.
- (31) Piecuch, P.; Paldus, J. *Int. J. Quantum Chem.* **1989**, *36*, 429–453.
- (32) Pal, S.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *160*, 212–218.
- (33) Mukhopadhyay, A.; Moitra, R. K.; Mukherjee, D. *J. Phys. B* **1979**, *12*, 1–18.
- (34) Mukherjee, D.; Pal, S. *Adv. Quantum Chem.* **1989**, *20*, 291–373.
- (35) Meissner, L.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *94*, 6670–6676.
- (36) Kaldor, U. *Theor. Chim. Acta* **1991**, *80*, 427–439.
- (37) Stolarczyk, L. Z.; Monkhorst, H. J. *Phys. Rev. A* **1985**, *32*, 725–742.
- (38) Meissner, L. *J. Chem. Phys.* **1995**, *103*, 8014–8021.
- (39) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2004**, *121*, 1670–1675.
- (40) Lindgren, I.; Mukherjee, D. *Phys. Rep.* **1987**, *151*, 93–127.
- (41) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2011**, *134*, 034106–1–12.
- (42) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2008**, *129*, 134105–1–12.
- (43) Hirao, K.; Nakatsuji, H. *J. Comput. Phys.* **1982**, *45*, 246–254.
- (44) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87–94.
- (45) Noga, J.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 7041–7050.
- (46) Noga, J.; Bartlett, R. J. *J. Chem. Phys.* **1988**, *89*, 3401.
- (47) Noga, J.; Bartlett, R. J.; Urban, M. *Chem. Phys. Lett.* **1987**, *134*, 126–132.
- (48) Casanova, D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *129*, 064104–1–12.
- (49) Dunning, T. K., Jr. *J. Chem. Phys.* **1970**, *53*, 2823.
- (50) Dunning, T. K., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (51) Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995–2016.
- (52) ACES II program is a product of the Quantum Theory Project, University of Florida. Authors: Stanton, J. F.; Gauss, J.; Watts, J. D.; Nooijen, M.; Oliphant, N.; Perera, S. A.; Szalay, P. G.; Lauderdale, W. J.; Kucharski, S. A.; Gwaltney, S. R.; Beck, S.; A. Balková, Musial, M.; Bernholdt, D. E.; Baeck, K. -K.; Sekino, H.; Rozyczko, P.; Huber, C.; Bartlett, R. J. Integral packages included are VMOL (Almlöf, J.; Taylor, P.), VPROPS (Taylor, P. R.), and a modified version of the ABACUS integral derivative package (Helgaker, T. U.; Jensen, H. J. Aa.; Olsen, J.; Jorgensen, P.; Taylor, P. R.).
- (53) Moore, C. E. Atomic Energy Levels. Natl. Bur. Stand. (U.S.), Circ.1949, Vol. 467.

Linear Scaling Hierarchical Integration Scheme for the Exchange-Correlation Term in Molecular and Periodic Systems

Asbjörn M. Burow and Marek Sierka*

Institut für Chemie, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany

S Supporting Information

ABSTRACT: An adaptive numerical integration scheme for efficient evaluation of the exchange-correlation term using localized basis functions and atom-centered grids is presented. The method treats molecules and systems with periodic boundary conditions on an equal footing. Its computational efficiency and $O(N)$ scaling with the system size is achieved by a hierarchical spatial grouping of basis functions and grid points using an octree. This allows for an efficient screening of negligible contributions and an optimal use of hardware-optimized matrix–matrix multiplication subroutines, such as BLAS. The implementation of the method within the TURBOMOLE program package demonstrates consistent accuracy and efficiency across molecular and periodic systems.

1. INTRODUCTION

Density functional theory (DFT) is distinguished from traditional wave function based *ab initio* methods by its treatment of the exchange-correlation contributions to the total energy as one-electron integrals, notwithstanding the fact that they originate from many-electron interactions. These one-electron integrals contain the exchange-correlation function^{1–9} $f(\mathbf{r})$ and are invariably too complex to permit their evaluation in analytical form. Consequently, the function $f(\mathbf{r})$ is integrated numerically on a finite set of grid points \mathbf{r}_m i.e.,

$$\int f(\mathbf{r}) \, d\mathbf{r} = \sum_m w_m \times f(\mathbf{r}_m) \quad (1)$$

with weights w_m depending on the grid point distribution. The strongly peaked nature of function f about nuclear centers requires special treatment and has inspired sophisticated algorithms, yielding numerical integration schemes with an optimum number of grid points per atom. There are different methodologies for the numerical evaluation implemented in various programs, such as DMOL,¹⁰ GAUSSIAN,¹¹ CRYSTAL,¹² ADF,¹³ MONDOSCF,¹⁴ deMon2k,¹⁵ Q-Chem,¹⁶ ORCA,¹⁷ and FHI-aims.¹⁸ For DFT calculations employing plane wave basis sets, uniformly spaced Cartesian grids are mainly used since they significantly simplify the fast Fourier transform. However, due to the strong variations of $f(\mathbf{r})$, these grids require special modification, e.g., the use of curvilinear coordinates,^{19–21} in order to obtain high accuracy. In addition, for DFT calculations using local basis sets, there are other ways to obtain grids adapted to the strongly peaked regions of $f(\mathbf{r})$ in polyatomic systems, e.g., hierarchical Cartesian grids¹⁴ and multicenter grids. The multicenter grids are composed of atom-centered grids and are restricted to the atoms of the unit cell (UC) in the case of periodic systems. The grid points of each atom-centered grid are usually arranged in Lebedev spheres^{22–24} of several radii.^{25–29} For such multicenter grids, several schemes^{10,11,30–36} are available to calculate the weights w_m of grid points \mathbf{r}_m belonging to the atoms i . A technique often applied is the renormalization of

atomic weights w_m^{at} using the continuous partition function $P_i(\mathbf{r})$, i.e.,

$$w_m = P_i(\mathbf{r}_m) \times w_m^{\text{at}} \quad (2)$$

There are several schemes to compute this partition function using, e.g., spherical atomic electron densities^{10,35} or purely geometric considerations.^{11,34} The atomic weights w_m^{at} i.e., the weights of isolated atoms, are usually determined using radial and spherical quadrature schemes.^{24–29,37}

For the multicenter grids, a highly efficient evaluation of the exchange-correlation term is achieved using hardware-optimized matrix–matrix multiplication subroutines, such as basic linear algebra subprograms (BLAS).^{38,39} For this, the necessary matrices are constructed by grouping grid points and basis functions into batches. Using local basis functions and prescreening techniques, each matrix contains exclusively the basis functions which have non-negligible contributions to the corresponding grid points. The multiplication of these matrices with blocks of the density matrix is required for the computation of the electron density and its gradients on the grid. For efficient use of the matrix–matrix multiplication subroutines, the dimensions of the matrices, i.e., the number of grid points and basis functions in the batches, should be as large as possible. However, local basis functions have non-negligible contributions only to a limited number of grid points. Therefore, the optimum choice of batch sizes has to balance the computational gain due to the efficiency of the matrix–matrix multiplications and the computational loss due to zero entries in the matrices caused by the local character of the basis functions. Although the use of batches is mentioned in several works,^{17,28,40–42} there are only a few that discuss batch shapes¹¹ and the use of hierarchical batching schemes.^{14,43} Hierarchical batching ideas were also used to add the exact and screened Fock exchange pieces in DFT.⁴⁴

Here, we present a new hierarchical scheme for efficient numerical evaluation of the exchange-correlation term within

Received: June 17, 2011

Published: August 08, 2011

orbital-based DFT employing localized basis functions. The method treats molecular and periodic systems of any dimensionality on an equal footing. It uses grid point batches of different sizes which overlap in space so that optimal batch sizes can be selected for different spatial extents of the basis functions. This is facilitated by grouping the grid points in batches belonging to the domains of an octree. Similar hierarchical structures are employed, e.g., for evaluation of the Coulomb term using the fast multipole method.^{45,46} This way, our present hierarchical numerical integration algorithm efficiently uses a multilevel batch approach combined with hardware-optimized matrix–matrix multiplications. Its implementation within the TURBOMOLE program package^{47,48} shows linear scaling with an increasing number of basis functions for a variety of molecular and periodic systems.

2. METHOD

2.1. Kohn–Sham Equations. In the orbital-based DFT under periodic boundary conditions (PBC), the canonical orbital coefficients matrix $\mathbf{C}^{\mathbf{k}}$ at each reciprocal lattice point \mathbf{k} satisfies

$$\mathbf{F}^{\mathbf{k}}\mathbf{C}^{\mathbf{k}} = \mathbf{S}^{\mathbf{k}}\mathbf{C}^{\mathbf{k}}\boldsymbol{\epsilon}^{\mathbf{k}} \quad (3)$$

with the Fock matrix $\mathbf{F}^{\mathbf{k}}$, the overlap matrix $\mathbf{S}^{\mathbf{k}}$, and the diagonal matrix $\boldsymbol{\epsilon}^{\mathbf{k}}$ of orbital energies. The elements $c_{\mu l}^{\mathbf{k}}$ of the matrix $\mathbf{C}^{\mathbf{k}}$ belong to the orbitals

$$\psi_i^{\mathbf{k}} = \sum_{\mu} c_{\mu i}^{\mathbf{k}} \phi_{\mu}^{\mathbf{k}} \quad (4)$$

constructed from periodic Bloch functions

$$\phi_{\mu}^{\mathbf{k}} = \sum_{\mathbf{L}} e^{i\mathbf{k}\cdot\mathbf{L}} \xi_{\mu}^{\mathbf{L}} \quad (5)$$

with local basis functions $\xi_{\mu}^{\mathbf{L}}$ centered at positions $\mathbf{R}_{\mu} + \mathbf{L}$, where \mathbf{R}_{μ} are the positions of the atoms in the UC and \mathbf{L} are the lattice vectors in direct space. The basis functions $\xi_{\mu}^{\mathbf{L}}$ are represented as linear combinations

$$\xi_{\mu}^{\mathbf{L}}(\mathbf{r}) = \sum_{\kappa=1}^{n_{\mu}} d_{\mu\kappa} \phi_{\mu\kappa}^{\mathbf{L}}(\mathbf{r}) \quad (6)$$

of a small number n_{μ} of primitive Cartesian Gaussian-type functions (CGTF) $\phi_{\mu\kappa}^{\mathbf{L}}$. The CGTF are defined as

$$\phi_{\mu\kappa}^{\mathbf{L}}(\mathbf{r}) = (x - R_{\mu x})^{\prime x} (y - R_{\mu y})^{\prime y} (z - R_{\mu z})^{\prime z} \times \exp(-\zeta_{\mu\kappa} |\mathbf{r} - \mathbf{R}_{\mu} - \mathbf{L}|^2) \quad (7)$$

with the Cartesian components x , y , and z of \mathbf{r} . The real exponent $\zeta_{\mu\kappa}$ and the integer exponents \prime_x , \prime_y , and \prime_z specify the radial and the angular parts of the function, respectively. Equations 3–7 hold also for the molecular case, where $\mathbf{k} = 0$ and $\mathbf{L} = 0$.

The Fock matrix $\mathbf{F}^{\mathbf{k}}$ is the sum of kinetic energy matrix $\mathbf{T}^{\mathbf{k}}$, Coulomb matrix $\mathbf{J}^{\mathbf{k}}$, and exchange–correlation matrix $\mathbf{X}^{\mathbf{k}}$. Using CGTF, the matrices $\mathbf{T}^{\mathbf{k}}$ and $\mathbf{J}^{\mathbf{k}}$ can be evaluated by analytical expressions, whereas the elements of $\mathbf{X}^{\mathbf{k}}$ are evaluated by numerical integration. The elements of the exchange–correlation matrix $\mathbf{X}^{\mathbf{k}}$ are defined as

$$X_{\mu\nu}^{\mathbf{k}} = \sum_{\mathbf{L}'} e^{i\mathbf{k}\cdot\mathbf{L}'} X_{\mu\nu}^{\mathbf{L}'} \quad \text{with } X_{\mu\nu}^{\mathbf{L}'} = \int \hat{O}_{\text{XC}}[\xi_{\mu}^{\mathbf{L}'} \xi_{\nu}^{\mathbf{L}'}] \, d\mathbf{r} \quad (8)$$

where $X_{\mu\nu}^{\mathbf{L}'}$ ($\mathbf{L}' = 0$ for molecules) and \hat{O}_{XC} are the real space exchange–correlation matrix and the exchange–correlation operator, respectively. For closed shell systems in the generalized gradient approximation, the operator \hat{O}_{XC} is expressed as

$$\hat{O}_{\text{XC}} = \frac{\partial f}{\partial \rho} + 2 \frac{\partial f}{\partial (|\nabla \rho|^2)} \nabla \rho \nabla \quad (9)$$

in terms of the exchange–correlation function^{1–9} $f(\rho(\mathbf{r}), |\nabla \rho(\mathbf{r})|^2)$ and the electron density ρ . The second term in eq 9 disappears in the local density approximation. For spin-dependent DFT in the unrestricted formalism, the form of \hat{O}_{XC} is presented elsewhere.^{1,49} The integration of the exchange–correlation function over the UC yields the exchange–correlation energy

$$E_{\text{XC}} = \int_{\text{UC}} f(\rho(\mathbf{r}), |\nabla \rho(\mathbf{r})|^2) \, d\mathbf{r} \quad (10)$$

for both closed shell and spin-unrestricted cases. For molecules, the integration in eq 10 spans the entire space.

2.2. Quadrature on the Grid. The terms $X_{\mu\nu}^{\mathbf{L}'}$ and E_{XC} in eqs 8 and 10 are evaluated numerically on a set of grid points \mathbf{r}_m belonging to the atoms i . The energy expression is given by

$$E_{\text{XC}} = \sum_i \sum_{m \in i} w_m \times f(\rho^m, |\nabla \rho^m|^2) \quad (11)$$

with the weights w_m defined in eq 2 and with $\rho^m = \rho(\mathbf{r}_m)$. In periodic systems^{50,51} and molecules, the summation runs over the atoms i of the UC and the whole molecule, respectively. The exchange–correlation matrix [eq 8] is evaluated numerically as

$$X_{\mu\nu}^{\mathbf{L}'} = \sum_i \sum_{m \in i} X_{\mu\nu}^{\mathbf{L}',m} \quad (12)$$

with contributions $X_{\mu\nu}^{\mathbf{L}',m}$ from the grid points \mathbf{r}_m defined as

$$X_{\mu\nu}^{\mathbf{L}',m} = w_m \sum_{\mathbf{L}} \hat{O}_{\text{XC}}[\xi_{\mu}^{\mathbf{L},m} \xi_{\nu}^{\mathbf{L}+\mathbf{L}',m}] \quad (13)$$

using weights w_m [eq 2]. Insertion of the closed shell form of the exchange–correlation operator \hat{O}_{XC} [eq 9] into eq 13 yields

$$X_{\mu\nu}^{\mathbf{L}',m} = \sum_{\mathbf{L}} (\xi_{\mu}^{\mathbf{L},m} z_{\nu}^{\mathbf{L}+\mathbf{L}',m} + z_{\mu}^{\mathbf{L},m} \xi_{\nu}^{\mathbf{L}+\mathbf{L}',m}) \quad (14)$$

with the potential

$$z_{\mu}^{\mathbf{L},m} = \frac{1}{2} F_{\mu}^m \xi_{\mu}^{\mathbf{L},m} + \sum_{\alpha} F_{\alpha}^m \xi_{\mu\alpha}^{\mathbf{L},m} \quad (15)$$

where $\xi_{\mu\alpha}^{\mathbf{L},m}$ and $\xi_{\mu\alpha}^{\mathbf{L},m} = \partial \xi_{\mu}^{\mathbf{L},m} / \partial \alpha$ ($\alpha = x, y, z$) denote values of the basis functions and their gradients, respectively, on the grid points \mathbf{r}_m . The terms F^m and F_{α}^m are defined as

$$F^m = w_m \frac{\partial f}{\partial \rho} \Big|_{\mathbf{r}_m} \quad (16)$$

$$(F_x^m, F_y^m, F_z^m) = 2w_m \frac{\partial f}{\partial (|\nabla \rho|^2)} \nabla \rho \Big|_{\mathbf{r}_m} \quad (17)$$

The molecular case is obtained setting $\mathbf{L} = 0$ and $\mathbf{L}' = 0$.

The computation of the exchange–correlation energy and the potential [eqs 11 and 15] requires evaluation of the electron density ρ and its gradient $\nabla \rho$ on the grid. The electron density ρ^m on the grid points \mathbf{r}_m is defined as the linear combination of basis

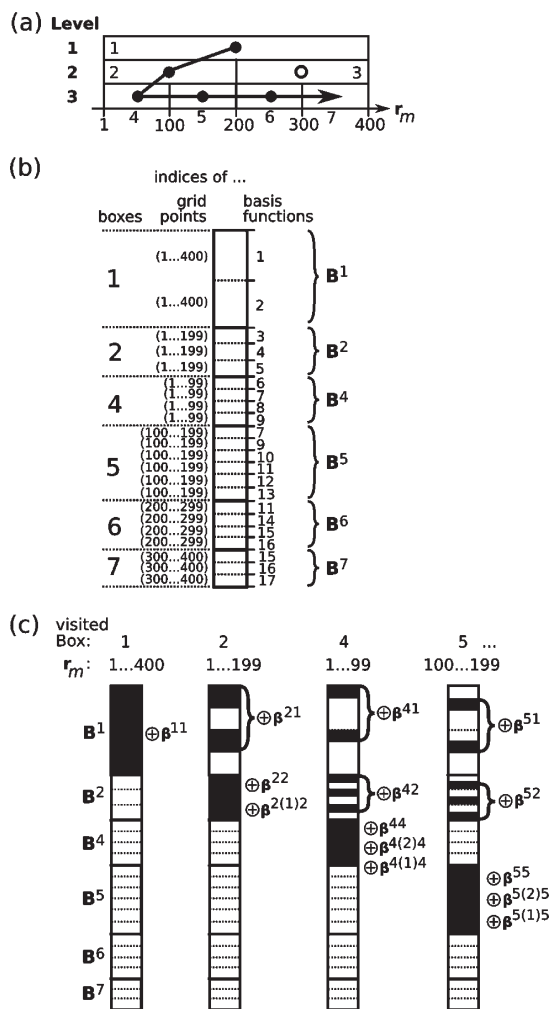


Figure 1. Example of the integration scheme. (a) Branch of boxes q in a binary tree (octree: eight children per parent box). Each box contains the grid points r_m . Box 3 is empty. (b) One-dimensional array of matrices B^q . Indices show grid points and basis functions assigned to boxes. (c) Accumulation (\oplus) of β^{pq} and $\beta^{(q)p}$ on B^q and B^p in loop 1 of Algorithm 1. For each visited box, the rectangle displays the array defined in b, where black areas highlight the changed content on the grid points r_m .

function products

$$\rho^m = \sum_{\mu L} \sum_{\nu L'} D_{\mu\nu}^{L'-L} \xi_{\mu\alpha}^{L,m} \xi_{\nu}^{L',m} \quad (18)$$

with the real space density matrix elements⁵² $D_{\mu\nu}^{L'-L}$ ($L = 0$ and $L' = 0$ for molecules). The partial derivatives of ρ with respect to the Cartesian coordinates $\alpha = x, y, z$ yield the components ρ_α of the gradient $\nabla\rho$, i.e.,

$$\rho_\alpha^m = \frac{\partial \rho}{\partial \alpha} \Big|_{r_m} = \sum_{\mu L} \sum_{\nu L'} D_{\mu\nu}^{L'-L} (\xi_{\mu\alpha}^{L,m} \xi_{\nu}^{L',m} + \xi_{\mu}^{L,m} \xi_{\nu\alpha}^{L',m}) \quad (19)$$

Since a limited number of local basis functions contributes to each grid point, the computational effort per grid point remains constant and the evaluation of the exchange-correlation term scales linearly with the system size (number of grid points).

2.3. Grouping Grid Points and Basis Functions. An efficient numerical computation of the electron density [eq 18] and the gradients [eq 19] on the grid is achieved using optimized matrix–matrix multiplication subroutines. For this, the necessary matrices are constructed by grouping grid points and basis functions using an octree. First, a cubic box enclosing all grid points is defined as the parent box of the octree. This parent box is bisected along each Cartesian axis to yield a set of eight child boxes belonging to the next lower level. Each of these child boxes is subdivided, creating children of the children, like the generations of a computational family tree. The number of divisions is chosen to yield box sizes optimal for the performance of the numerical integration. In the next step, all grid points and basis functions are sorted into the octree. Grid points are sorted into the boxes of the tree by coordinates. The sorting of basis functions into boxes is based on their spatial extents.^{11,53} For this, spheres of the radii⁵³

$$r_\mu = \max_{\kappa} \left(\sqrt{\frac{-\ln \varepsilon + 0.5 \times \ln \xi_{\mu\kappa}^L}{\xi_{\mu\kappa}^L}} \right) \quad (20)$$

centered at the atomic positions $\mathbf{R}_\mu + \mathbf{L}$ are assigned to the basis functions ξ_μ^L [eq 6], where ε is a precision parameter. Starting from the highest level, each basis function is assigned to the boxes at a given level which are fully enclosed by the corresponding sphere. If the basis function is assigned to a box, then its child boxes at the following levels are removed from the assignment for this function. Occupied boxes contain at least one grid point and one basis function. Root boxes are occupied boxes without any occupied parent at higher levels. A branch is formed by a root box and its occupied child boxes. The boxes p of a branch are ordered by levels from top to bottom (see Figure 1).

2.4. Exchange-Correlation Term in the Octree. This section describes the evaluation of the exchange-correlation term in the octree. First, the necessary terms and notations are defined. For each box p of the octree, the values of basis functions on the grid, $\xi_\mu^{L,m}$, are arranged in a rectangular $M_p \times K_p$ matrix ξ^p (i.e., a batch), where M_p and K_p are the number of grid points and the number of basis functions, respectively. Similarly, ξ_α^p denotes the matrices containing gradients of basis functions on the grid points in the box p . The $M_p \times K_q$ matrices $\xi^{p(q)}$ contain the values of the basis functions, $\xi_\mu^{L,m}$, assigned to a parent box q , on the grid points of its child box p . The $K_p \times K_q$ matrices D^{pq} denote submatrices of the density matrix. They are formed from the elements $D_{\mu\nu}^{L'-L}$ belonging to the functions ξ_μ^L and $\xi_\nu^{L'}$, assigned to the boxes p and q , respectively. The $M_q \times K_q$ matrices B^q defined for each box q are used to accumulate the $M_p \times K_q$ matrices

$$\beta^{pq} = \xi^p D^{pq} \quad (21)$$

calculated for child boxes p . This accumulation is denoted as

$$B^q \leftarrow B^q \oplus \beta^{pq} \quad (22)$$

The columns of the matrices B^q and β^{pq} correspond to the same set of K_q basis functions. The rows of B^q correspond to M_q grid points in a box q , and the rows of β^{pq} correspond to M_p grid points of a child box p . The m th row of a matrix \mathbf{M} is denoted as $\text{row}_m[\mathbf{M}]$. Using these definitions, Algorithm 1 shows the computation of the electron density ρ and its gradients ρ_α on the grid for a single branch of the octree. The electron density and its gradients in different branches are independent of each other.

Algorithm 1. Compute Electron Density and Gradients within a Single Branch

- (1) For each box p of the branch:
 - (2) Initialize \mathbf{B}^p to zero
 - (3) Compute and save: ξ^p and ξ_α^p
 - (4) For each box q with $q = (p$ and all parent boxes of $p)$:
 - (5) Form \mathbf{D}^{pq}
 - (6) $\beta^{pq} = \xi^p \mathbf{D}^{pq}$
 - (7) $\mathbf{B}^q \leftarrow \mathbf{B}^q \oplus \beta^{pq}$
 - (8) If $q \neq p$:
 - (9) $\beta^{p(q)p} = \xi^{p(q)} (\mathbf{D}^{pq})^T$
 - (10) $\mathbf{B}^p \leftarrow \mathbf{B}^p \oplus \beta^{p(q)p}$
- (11) Initialize ρ and ρ_α to zero
- (12) For each box p of the branch:
 - (13) For each \mathbf{r}_m in p :
 - (14) $\rho^m \leftarrow \rho^m + \text{row}_m[\mathbf{B}^p] \times \text{row}_m[\xi^p]^T$
 - (15) $\rho_\alpha^m \leftarrow \rho_\alpha^m + 2 \times \text{row}_m[\mathbf{B}^p] \times \text{row}_m[\xi_\alpha^p]^T$

Figure 1 shows an example for the storage of the matrices \mathbf{B}^p of a branch as a one-dimensional array and the accumulation of the matrices β^{pq} and $\beta^{p(q)p}$ in steps 7 and 10 of Algorithm 1. The values of the electron density and the gradients on the grid obtained with Algorithm 1 are used in the Algorithm 2, which describes the numerical integration for the exchange-correlation energy E_{XC} [eq 11] and the matrix elements $X_{\mu\nu}^L$ [eq 12].

Algorithm 2. Compute E_{XC} and Matrix Elements $X_{\mu\nu}^L$

- (1) For each branch:
 - (2) Perform Algorithm 1 (compute and save: $\xi^p, \xi_\alpha^p, \rho, \rho_\alpha$)
 - (3) {For each \mathbf{r}_m in the root box:
 - (4) $E_{\text{XC}} \leftarrow E_{\text{XC}} + w_m \times f(\rho^m, |\nabla \rho^m|^2)$
 - (5) Compute and save: F^m, F_x^m, F_y^m, F_z^m
 - (6) For each box p of the branch:
 - (7) {For each ξ_μ^L and \mathbf{r}_m in p :
 - (8) Compute and save: $z_\mu^{L,m}$
 - (9) For each box q with $q = (p$ and all parent boxes of $p)$:
 - (10) For each ξ_μ^L in p and ξ_ν^L in q :
 - (11) For each \mathbf{r}_m in p :
 - (12) $X_{\mu\nu}^{L-L} \leftarrow X_{\mu\nu}^{L-L} + \xi_\mu^{L,m} z_\nu^{L,m} + z_\mu^{L,m} \xi_\nu^{L,m}$
 - (13) If $q \neq p$: $X_{\nu\mu}^{L-L} \leftarrow X_{\nu\mu}^{L-L} + \xi_\nu^{L,m} z_\mu^{L,m} + z_\nu^{L,m} \xi_\mu^{L,m}$

Our implementation uses the symmetry

$$X_{\mu\nu}^{L-L} = X_{\nu\mu}^{L-L} \quad (23)$$

in steps 12 and 13 of Algorithm 2. The computational demand of steps 2, 3, and 6 of Algorithm 2 is independent of the system size, while the number of branches scales with $O(N)$. The computational efficiency is achieved using optimized matrix–matrix multiplications for products between the matrices ξ and \mathbf{D} in steps 6 and 9 of Algorithm 1. This becomes so efficient that the evaluation of the matrix elements in steps 12 and 13 of Algorithm 2 can be the bottleneck of the calculation. Therefore, a two-level screening procedure is applied for efficient evaluation of elements $X_{\mu\nu}^{L-L}$. First, an upper bound $X^{\max(pq)}$ of the contributions $X_{\mu\nu}^{L-L,m}$ from each pair of boxes p, q is determined within loop 9 of Algorithm 2 with

$$X^{\max(pq)} = \xi^{\max(p)} z^{\max(q)} + z^{\max(p)} \xi^{\max(q)} \quad (24)$$

employing the maximum values of basis functions, $\xi^{\max(s)}$, and potentials, $z^{\max(s)}$, in the boxes $s = p, q$,

Table 1. Specifications of Grids^a

grid	$n_\omega^{\text{outer},b}$	H, He		Li–Ne		Na–Ar	
		n_r	n_{tot}	n_r	n_{tot}	n_r	n_{tot}
3	302	10/5/15	5340	11/6/18	6382	13/7/20	7148
5	590	18/9/28	17978	20/10/30	19320	21/11/33	21226
7	1202	28/14/43	53954	30/15/45	56520	31/16/48	60262

^a Number of outer angular grid points, n_ω^{outer} ; number of radial grid points, n_r (split into $n_r^{\text{inner}}/n_r^{\text{medium}}/n_r^{\text{outer}}$); and total number of grid points per atom, n_{tot} . ^b The number of angular grid points in the inner and medium regions are always $n_\omega^{\text{inner}} = 26$ and $n_\omega^{\text{medium}} = 110$, respectively.

$$\xi^{\max(s)} = \max_{\{\mathbf{r}_m, \mu, \mathbf{L}\} \in s} (\xi_\mu^{L,m}) \quad (25)$$

$$z^{\max(s)} = \max_{\{\mathbf{r}_m, \mu, \mathbf{L}\} \in s} (z_\mu^{L,m}) \quad (26)$$

All contributions $X_{\mu\nu}^{L-L,m}$ from the pair of boxes p, q are neglected in steps 12 and 13 of Algorithm 2 if the upper bound $b = X^{\max(pq)}$ satisfies

$$b \times M_p < \tau \quad (27)$$

with a threshold parameter τ . Second, if eq 27 is not satisfied, then an upper bound $X_{\mu\nu}^{L-L, \max(pq)}$ of the contributions $X_{\mu\nu}^{L-L,m}$ from each pair of basis functions ξ_μ^L, ξ_ν^L within the boxes pair p, q is determined with

$$X_{\mu\nu}^{L-L, \max(pq)} = \xi_\mu^{L, \max(p)} z_\nu^{L, \max(q)} + z_\mu^{L, \max(p)} \xi_\nu^{L, \max(q)} \quad (28)$$

For this, the maximum values of basis functions and potentials for each pair μ, \mathbf{L} in a box $s = p, q$ are used:

$$\xi_\mu^{L, \max(s)} = \max_{\mathbf{r}_m \in s} (\xi_\mu^{L,m}) \quad (29)$$

$$z_\mu^{L, \max(s)} = \max_{\mathbf{r}_m \in s} (z_\mu^{L,m}) \quad (30)$$

If eq 27 is satisfied for $b = X_{\mu\nu}^{L-L, \max(pq)}$, then the contributions $X_{\mu\nu}^{L-L,m}$ of the corresponding pair of basis functions are neglected for the boxes pair p, q .

3. COMPUTATIONS

The numerical accuracy and the scaling of our method are investigated using several molecular and periodic systems. The errors in the calculated electron numbers and the convergence of exchange-correlation energies with increasing grid size are used as criteria for assessment of the numerical accuracy. Electron numbers and exchange-correlation energies are obtained by numerical integration using converged electron densities from self-consistent field (SCF) calculations with a convergence threshold of 1.0×10^{-10} hartree for total energy and exchange-correlation energy. The scaling of the numerical integration with respect to increasing system sizes is investigated measuring CPU timings for a single SCF step.

3.1. Computational Details. All calculations are performed using the B-LYP exchange-correlation functional^{6,7} with double-, triple-, and quadruple- ζ valence split basis sets plus polarization

functions (DZVP,⁵⁴ TZVP,⁵⁵ and QZVP,⁵⁶ respectively) and standard grid sizes²⁸ 3, 5, and 7 of TURBOMOLE. The details of the integration grids are shown in Table 1. The total numbers of grid points per atom, n_{tot} can be calculated as²⁸

$$n_{\text{tot}} = \sum_{\alpha} n_{\omega}^{\alpha} \times n_r^{\alpha} \text{ with } \alpha = \text{inner, medium, outer} \quad (31)$$

with the numbers of angular (n_{ω}^{α}) and radial (n_r^{α}) grid points of each atomic grid region α . The n_{ω}^{α} angular points of each atom form the angular grid on a unit sphere. The angular grids in Table 1 contain 26, 110, 302, 590, and 1202 grid points. These grids have been developed by Lebedev et al.^{22–24} for accurate numerical integration of spherical harmonics $Y_{l \leq L, m}$ with $L = 7, 17, 29, 41,$ and 59 . It should be noted that the numbers of points in Table 1 belong to the current TURBOMOLE version.⁴⁷ For grids 3 and 5, these numbers are larger than grid sizes 3 and 5 of the first implementation (ref 28).

Our implementation of the (normalized) partition functions $P_i(\mathbf{r})$ [eq 2] follows the ideas of Becke³⁴ and uses the form

$$P_i(\mathbf{r}) \equiv P_i^0(\mathbf{r}) = \tilde{P}_i^0(\mathbf{r}) / \sum_{j\mathbf{L}} \tilde{P}_j^{\mathbf{L}}(\mathbf{r}) \quad (32)$$

with indices i and j of atoms in the reference unit cell and lattice vectors \mathbf{L} . The unnormalized partition functions $\tilde{P}_j^{\mathbf{L}}$ are defined as

$$\tilde{P}_j^{\mathbf{L}}(\mathbf{r}) = \prod_{k\mathbf{L}' \neq j\mathbf{L}} s_{jk}^{\mathbf{L}\mathbf{L}'}(\mathbf{r}) \quad (33)$$

where indices k and \mathbf{L}' run over all atoms and lattice vectors, respectively, with exception of the index pair $j\mathbf{L}$. The scaled step functions

$$s_{jk}^{\mathbf{L}\mathbf{L}'}(\mathbf{r}) = 0.5 \times [1 - h_{jk}^{\mathbf{L}\mathbf{L}'}(\mathbf{r})] \quad (34)$$

are formed using the polynomial step functions $h_{jk}^{\mathbf{L}\mathbf{L}'}$ as defined in ref 11 (cf. eq 14 in this reference), i.e.,

$$h_{jk}^{\mathbf{L}\mathbf{L}'} = \frac{1}{16} \times [35\nu_{jk}^{\mathbf{L}\mathbf{L}'} - 35(\nu_{jk}^{\mathbf{L}\mathbf{L}'})^3 + 21(\nu_{jk}^{\mathbf{L}\mathbf{L}'})^5 - 5(\nu_{jk}^{\mathbf{L}\mathbf{L}'})^7], \nu_{jk}^{\mathbf{L}\mathbf{L}'} \in [-1, 1] \quad (35)$$

with the modified confocal elliptical coordinates $\nu_{jk}^{\mathbf{L}\mathbf{L}'}$

$$\nu_{jk}^{\mathbf{L}\mathbf{L}'}(\mathbf{r}) = \mu_{jk}^{\mathbf{L}\mathbf{L}'}(\mathbf{r})/a, a \in]0, 1] \quad (36)$$

The relative half-step width a is set to 0.64 as suggested in ref 11. The confocal elliptical coordinates $\mu_{jk}^{\mathbf{L}\mathbf{L}'}$ have the form

$$\mu_{jk}^{\mathbf{L}\mathbf{L}'}(\mathbf{r}) = \frac{|\mathbf{r} - \mathbf{R}_j - \mathbf{L}| - |\mathbf{r} - \mathbf{R}_k - \mathbf{L}'|}{|\mathbf{R}_j + \mathbf{L} - \mathbf{R}_k - \mathbf{L}'|} \quad (37)$$

with position vectors \mathbf{R}_j and \mathbf{R}_k of atoms j and k , respectively, in the reference unit cell. In cases where $\mu_{jk}^{\mathbf{L}\mathbf{L}'}$ results in $\nu_{jk}^{\mathbf{L}\mathbf{L}'} \notin [-1, 1]$, the values of the step function $h_{jk}^{\mathbf{L}\mathbf{L}'}$ are set either to 0 (for $\nu_{jk}^{\mathbf{L}\mathbf{L}'} < -1$) or to 1 (for $\nu_{jk}^{\mathbf{L}\mathbf{L}'} > 1$). For the molecular case, only the terms with $\mathbf{L} = \mathbf{L}' = 0$ remain in eqs 32–37. A fixed cutoff radius r_{cut} similar to the one suggested in ref 50 is used for the truncation of the infinite sums and products in eqs 32 and 33. If the distance between the atoms at positions \mathbf{R}_j and $\mathbf{R}_j + \mathbf{L}$ is larger than r_{cut} then the contribution $\tilde{P}_{j\mathbf{L}}$ is neglected in eq 32. Similarly, the contributions $s_{jk}^{\mathbf{L}\mathbf{L}'}$ are neglected in eq 33 for atoms at positions $\mathbf{R}_j + \mathbf{L}$ and $\mathbf{R}_k + \mathbf{L}'$ with distances larger than r_{cut} . Cutoff radii of 28 and 39 bohr

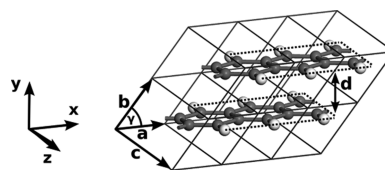


Figure 2. The 3×2 supercell of the aromatic chains ($d = 2.451 \text{ \AA}$). Dark and light spheres represent C and H atoms, respectively. The periodic direction of 1D and 1D (aper.) models belongs to cell parameter a . The periodic directions of 2D models correspond to cell parameters a and b .

Table 2. Dimensions m and n of $m \times n$ Supercells for Aromatic Chains^a

model	m	n
1D	1, 2, 3, 6, 12, 23, 46, 92	1
1D (aper.) ^b	1	1, 2, 4, 8, 16, 32
2D	1, 2	1
	2, 3	2
	3, 4	3
	4	4

^aThe periodic direction of 1D and 1D (aper.) models corresponds to cell parameter a (dimension m). For 2D models, the dimensions m and n correspond to the periodic directions of cell parameters a and b , respectively. ^b See text for explanation.

are used for grids 3 and 5 and grid 7, respectively. These cutoffs yield converged values for the calculated numbers of electrons.

The precision parameter ε [eq 20] and the threshold parameter τ [eq 27] are both set to 1×10^{-9} a.u. The program is compiled with the Portland Group Fortran 95 (PGF95) compiler version 8.0, and calculations are performed on a single x86–64 CPU (Quad-Core AMD Opteron Processor 2354, 2211.377 MHz, 512 KB cache size) under the SUSE Linux operating system using kernel version 2.6.16.

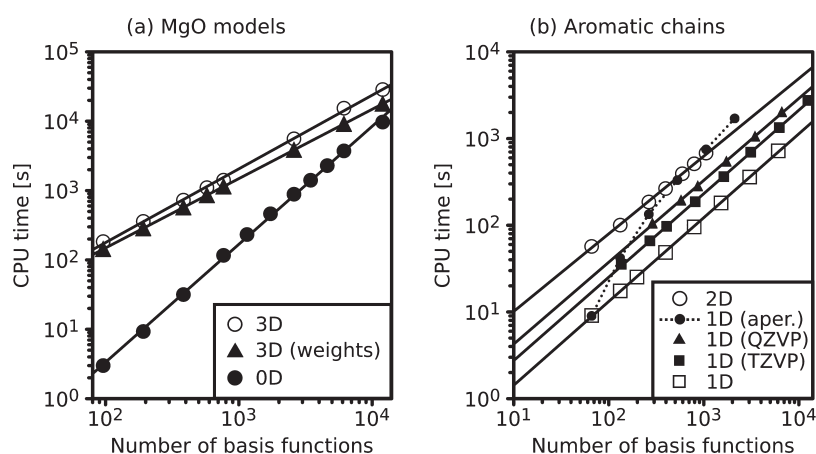
3.2. Models. The computational scaling of the algorithm is investigated using series of bulk MgO pieces, both molecular and 3D periodic, as well as 1D and 2D periodic models of aromatic chains. The MgO models are constructed from the conventional cell (lattice constant 4.211 Å, $Fm\bar{3}m$ space group) containing four MgO units and range from $1 \times 1 \times 1$ to $5 \times 5 \times 5$ supercells. For 1D and 2D models of aromatic chains, condensed benzene rings (bond distances $b(\text{C}–\text{C}) = 1.530 \text{ \AA}$ and $b(\text{C}–\text{H}) = 1.179 \text{ \AA}$) are used. The constructed 1×1 UC with dimensions $a = 2.65 \text{ \AA}$, $b = 3.20 \text{ \AA}$, and $\gamma = 50^\circ$ has a C_4H_2 composition (Figure 2). Table 2 shows dimensions of the supercells used. Models denoted “1D” and “1D (aper.)” use a series of supercells increasing along the periodic direction (corresponding to a) and aperiodic direction (corresponding to b), respectively. Table 3 presents the systems used to assess the accuracy of the method. These systems range from zero dimensional (0D) molecules over periodic chains (1D) and surfaces (2D) to bulk structures (3D; see Supporting Information).

3.3. Results. Table 3 shows the relative errors of calculated electron numbers, $|\Delta N/N|$, and the exchange-correlation energies E_{XC} [eq 10]. The magnitudes of $|\Delta N/N|$ decrease from 4×10^{-6} for grid 3 to 3×10^{-8} for grid 7 and are comparable to errors obtained with previous integration schemes.²⁸ The exchange-correlation energies of grids 3 and 5 deviate from the E_{XC}

Table 3. Relative Errors of the Number of Electrons, $|\Delta N/N|$, and Exchange-Correlation Energies, E_{XC} (hartree)^a

model (periodicity)	$ \Delta N/N $			ΔE_{XC}		E_{XC}
	grid					
	3	5	7	3	5	7
H ₂ O (0D) ^b	1.3D-6	2.4D-8	4.9D-10	9.3D-6	-1.3D-7	-9.30719714
CH ₄ (0D) ^b	1.3D-6	2.3D-8	7.0D-10	-1.0D-5	2.0D-7	-6.86699820
C ₈ H ₁₈ (0D) ^b	6.9D-6	1.1D-7	1.2D-8	9.8D-5	2.0D-6	-49.59078971
SiH ₄ (0D) ^b	8.6D-6	3.0D-7	1.6D-8	1.5D-5	-7.4D-7	-22.36565262
SiH ₄ (0D) ^c	9.4D-6	3.2D-7	1.6D-8	-4.9D-7	-3.6D-7	-22.34328264
[C ₂ H ₅ OH] ₂ (0D) ^c	6.4D-7	5.6D-8	3.5D-10	1.7D-5	4.9D-7	-43.91399349
[C ₂ H ₅ OH] ₂ (1D) ^c	7.2D-7	4.2D-8	2.8D-9	1.6D-5	6.3D-7	-43.90002316
[C ₂ H ₅ OH] ₂ (2D) ^c	1.6D-6	7.4D-8	3.8D-8	2.2D-5	-3.4D-9	-43.89965668
[C ₂ H ₅ OH] ₂ (3D) ^c	1.4D-6	5.7D-7	8.1D-8	2.5D-5	-3.0D-6	-43.89806291
[(H ₂ N) ₂ CO] ₂ (3D) ^c	1.1D-5	2.1D-6	5.8D-8	-3.1D-5	4.1D-6	-58.58157643
(SiO ₂) ₃ (3D) ^{c,d}	5.6D-6	7.4D-7	5.9D-8	1.7D-4	3.4D-5	-115.00032197
$ \Delta $ ^e	3.7D-6	2.1D-7	2.6D-8	3.8D-5	4.1D-6	

^a E_{XC} of grids 3 and 5 are given by $E_{XC}(\text{grid } 7) + \Delta E_{XC}$. ^b Calculated with TZVP basis sets. ^c Calculated with DZVP basis sets. ^d β -quartz. ^e Average of unsigned deviations

Figure 3. CPU timings of E_{XC} and matrix elements $X_{\mu\nu}^L$ in one SCF step (DZVP basis sets and grid 5, if not stated otherwise).Table 4. Scaling $O(N^x)$ of the Numerical Integration (Energy E_{XC} and Matrix $X_{\mu\nu}^L$) with the System Size and Relative Errors of Electron Numbers, $|\Delta N/N|$, Using Grid 5

model	periodicity	basis	x	$ \Delta N/N $
(Mg ₄ O ₄) _n	0D	DZVP	1.7	4.6D-7
	3D	DZVP	1.1	6.1D-7
(C ₄ H ₂) _n	1D	DZVP	1.0	2.4D-7
	1D	TZVP	1.0	3.6D-7
	1D	QZVP	0.9	2.8D-7
	1D (aper.)	DZVP	1.2 ^a	3.0D-7
	2D	DZVP	0.9	1.4D-6

^a For the supercells 1×8 , 1×16 , and 1×32 .

of grid 7 in the range of 4×10^{-5} and 4×10^{-6} hartree, respectively. This demonstrates the consistent evaluation of the integrals on different grids. The use of better grids indicates the convergence of the numerically calculated values toward the exact integrals.

The CPU times required for the numerical evaluation of E_{XC} and the matrix elements $X_{\mu\nu}^L$ [eq 8] are shown in Figure 3. The fitted scaling factors x summarized in Table 4 demonstrate linear scaling $O(N^{x \approx 1.0})$ of our algorithm except for the series of models with increasing size in aperiodic directions, i.e., the 1D (aper.) aromatic chains and the magnesium oxide clusters. For the 1D (aper.) aromatic chains, the scaling factor $x = 1.2$ is obtained for larger supercells (1×8 , 1×16 , and 1×32 ; cf. Figure 3). For magnesium oxide clusters, linear scaling is not achieved ($x = 1.7$) within the investigated system sizes. In general, the size of the basis set (DZVP, TZVP, and QZVP) influences only the prefactor and not the scaling of the method.

For the bulk MgO models, the computation of the weights [eq 2] scales perfectly linearly but is almost as time-demanding as the exchange-correlation term itself (Figure 3). However, the weights have to be computed only once for a given structure, whereas the exchange-correlation term has to be computed for each SCF step. The $5 \times 5 \times 5$ supercell of bulk MgO contains 12 000 basis functions. For this size, the computation of the weights and the exchange-correlation term requires 5 and 8 h, respectively.

Table 5. Total Number of Contributing Grid Points, n_{pt} , the Average Number of Non-Negligible Basis Functions Contributions Per Grid Point, n_{fpp} and the Average CPU Time T (μs) for Evaluation of One Basis Function Value on a Grid Point for Selected MgO and Aromatic Chain Models (DZVP Basis Sets and Grid 5, if Not Stated Otherwise)

model	n_{pt}	n_{fpp}	T	n_{pt}	n_{fpp}	T
(Mg ₄ O ₄) _n		0D			3D	
1 × 1 × 1	1.4D+5	32	0.6	1.2D+5	871	1.8
2 × 2 × 2	1.0D+6	122	0.9	9.3D+5	781	1.5
3 × 3 × 3	3.4D+6	228	1.1	3.1D+6	754	2.4
4 × 4 × 4	7.9D+6	315	1.5	7.4D+6	741	2.8
5 × 5 × 5	1.5D+7	377	1.7	1.5D+7	733	2.7
(C ₂ H ₄) _n		1D			1D(QZVP)	
1 × 1	9.6D+4	81	1.2	9.6D+4	187	5.8
6 × 1	5.8D+5	75	1.1	5.8D+5	171	5.5
12 × 1	1.2D+6	74	1.1	1.2D+6	167	5.5
23 × 1	2.2D+6	74	1.1	2.2D+6	168	5.4
(C ₂ H ₄) _n		1D (aper.)			2D	
1 × 1	9.6D+4	81	1.2	8.6D+4	294	2.3
1 × 4 ^a	3.5D+5	204	1.8	3.4D+5	268	2.0
1 × 16 ^b	1.4D+6	265	2.0	7.7D+5	259	2.0
1 × 32 ^c	2.8D+6	287	2.2	1.4D+6	254	2.0

^a Supercell for 2D: 2 × 2. ^b Supercell for 2D: 3 × 3. ^c Supercell for 2D: 4 × 4.

The total number of contributing grid points, n_{pt} , and the average number of non-negligible basis functions contributions per grid point, n_{fpp} , for selected MgO and aromatic chain models are shown in Table 5. These values are used to estimate the average time T for evaluation of one basis function value on a grid point (Table 5)

$$T = \frac{t_{\text{CPU}}}{n_{\text{pt}} \times n_{\text{fpp}}} \quad (38)$$

Calculations for bulk MgO and for aromatic chain models using the QZVP basis sets yield $T \approx 2$ and $T \approx 6 \mu\text{s}$, respectively. Therefore, the increased basis set size has a pronounced effect on T . This is because extended basis sets lead to a larger total number of non-negligible contributions. In remaining cases, the values of T are below $2 \mu\text{s}$. The values of T facilitate future comparisons of the performance of our method with other numerical integration schemes.

4. COMPARISON WITH EXISTING METHODS

The idea of employing hierarchical structures such as the octree in integration schemes for DFT has also been explored by other authors. For example, Havu et al.⁴³ have investigated the application of different hierarchical structures, including the octree, in numerical integrations for both molecular and periodic systems. In their method, the levels of a tree are used to select appropriate box sizes for different spatial regions, depending on the distribution of the grid points and basis functions. However, different boxes do not overlap in space and for numerical evaluation of integrals all boxes are merged into one level. In contrast, our scheme uses boxes of different sizes at different octree levels which do overlap in space. In fact, this overlap is

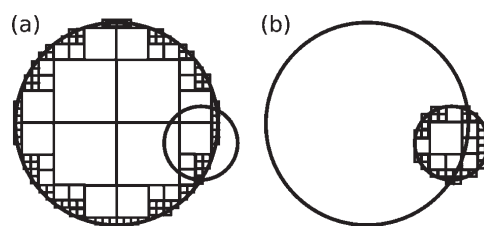


Figure 4. Schematic representation of two overlapping basis functions with different extents (shown as circles) sorted in the levels of an octree (boxes of the octree are shown as squares). The integration of the more diffuse function uses different box sizes than the integration of the compact one. This requires overlapping sets of boxes at different octree levels for the same spatial region.

crucial for the efficiency of the method. To illustrate this, consider two overlapping local basis functions with very different extents, as shown in Figure 4. For each basis function, the integration boxes are larger close to their centers and decrease in size at their boundaries. Since both basis functions overlap in space the numerical integration can only be efficient if boxes of different sizes are used within the same spatial region, as in our method. If just one box is selected for each region, like in the method of Havu et al., then the size of this box may not be optimal to evaluate the contributions of both functions efficiently. In a different approach, Challacombe¹⁴ used a tree structure to construct adaptive Cartesian grids. In contrast, our method is more general and can be applied to arbitrary integration grids and any type of local basis functions. This facilitates implementation of the method in existing DFT programs.

5. SUMMARY

An adaptive numerical integration scheme for efficient evaluation of the exchange correlation term using localized basis functions is presented. The method treats molecular and periodic systems on an equal footing. Its computational efficiency and $O(N)$ scaling with the system size is achieved by hierarchical spatial grouping of grid points and basis functions using an octree. This grouping is used to form matrices from the values of basis functions on the grid points and allows for an optimal use of hardware-optimized matrix–matrix multiplication subroutines, such as BLAS. The novel aspect of our method is the multilevel approach using matrices of different sizes in the same region of space. This leads to the evaluation of exchange–correlation contributions with matrices of optimal sizes for basis functions of various spatial extents. The optimum choice of the matrix sizes balances the computational gain due to the efficiency of matrix–matrix multiplications and the computational loss due to negligible entries in the matrices caused by the local character of the basis functions. The implementation of the method within the TURBOMOLE program package shows linear scaling for a variety of molecular and periodic systems. The required CPU time to evaluate one basis function value on a grid point is around $2 \mu\text{s}$ for most of the systems calculated with the DZVP basis sets.

■ ASSOCIATED CONTENT

S Supporting Information. Coordinates of structures listed in Table 3. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: marek.sierka@chemie.hu-berlin.de.

ACKNOWLEDGMENT

This work was supported by the Deutsche Forschungsgemeinschaft (Center of Excellence UNICAT and Sonderforschungsbereich 546), the Stiftung Stipendien-Fonds des Verbandes der Chemischen Industrie, and TURBOMOLE GmbH. The authors are grateful to Professor Filipp Furche for stimulating discussions and to Professor Joachim Sauer for support.

REFERENCES

- (1) Johnson, B. G.; Gill, P. M. W.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 5612–5626.
- (2) Herman, F.; Van Dyke, J. P.; Ortenburger, I. B. *Phys. Rev. Lett.* **1969**, *22*, 807–811.
- (3) Perdew, J. P.; Yue, W. *Phys. Rev. B* **1986**, *33*, 8800–8802.
- (4) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (5) Ghosh, S. K.; Parr, R. G. *Phys. Rev. A* **1986**, *34*, 785–791.
- (6) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (7) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (8) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (9) Cedillo, A.; Robles, J.; Gázquez, J. L. *Phys. Rev. A* **1988**, *38*, 1697–1701.
- (10) Delley, B. *J. Chem. Phys.* **1990**, *92*, 508–517.
- (11) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *257*, 213–223.
- (12) Dovesi, R.; Saunders, V. R.; Roetti, R.; Orlando, R.; Zicovich-Wilson, C. M.; Pascale, F.; Civalleri, B.; Doll, K.; Harrison, N. M.; Bush, I. J.; D'Arco, P.; Llunell, M. *CRYSTAL09 User's Manual*; University of Torino: Torino, Italy, 2009.
- (13) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.
- (14) Challacombe, M. *J. Chem. Phys.* **2000**, *113*, 10037–10043.
- (15) Geudtner, G.; Janetzko, F.; Köster, A. M.; Vela, A.; Calaminici, P. *J. Comput. Chem.* **2006**, *27*, 483–490.
- (16) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.
- (17) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. *Chem. Phys.* **2009**, *356*, 98–109.
- (18) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.
- (19) Gygi, F. *Phys. Rev. B* **1993**, *48*, 11692–11700.
- (20) Pérez-Jordá, J. M. *Phys. Rev. A* **1995**, *52*, 2778–2784.
- (21) Rodríguez, J. I.; Thompson, D. C.; Ayers, P. W.; Köster, A. M. *J. Chem. Phys.* **2008**, *128*, 224103.
- (22) Lebedev, V. I. *Siberian Math. J.* **1977**, *18*, 99–107.
- (23) Delley, B. *J. Comput. Chem.* **1996**, *17*, 1152–1155.
- (24) Lebedev, V. I.; Laikov, D. N. *Doklady Math.* **1999**, *59*, 477–481.
- (25) Gill, P. M. W.; Johnson, B. G.; Pople, J. A. *Chem. Phys. Lett.* **1993**, *209*, 506–512.
- (26) Murray, C. W.; Handy, N. C.; Laming, G. J. *Mol. Phys.* **1993**, *78*, 997–1014.
- (27) Baker, J.; Andzelm, J.; Scheiner, A.; Delley, B. *J. Chem. Phys.* **1994**, *101*, 8894–8902.
- (28) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346–354.
- (29) Mura, M. E.; Knowles, P. J. *J. Chem. Phys.* **1996**, *104*, 9848–9858.
- (30) Boerrigter, P. M.; te Velde, G.; Baerends, E. J. *Int. J. Quantum Chem.* **1988**, *33*, 87–113.
- (31) Averill, F. W.; Painter, G. S. *Phys. Rev. B* **1989**, *39*, 8115–8121.
- (32) te Velde, G.; Baerends, E. J. *J. Comput. Phys.* **1992**, *99*, 84–98.
- (33) Pederson, M. R.; Jackson, K. A. *Phys. Rev. B* **1990**, *41*, 7453–7461.
- (34) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 2547–2553.
- (35) Savin, A. *Int. J. Quantum Chem.* **1988**, *34* (S22), 59–69.
- (36) Lin, Z.; Jaffe, J. E.; Hess, A. C. *J. Phys. Chem. A* **1999**, *103*, 2117–2127.
- (37) Lindh, R.; Malmqvist, P.-Å.; Gagliardi, L. *Theor. Chem. Acc.* **2001**, *106*, 178–187.
- (38) BLAS (Basic Linear Algebra Subprograms). <http://www.netlib.org/blas> (accessed July 25, 2011).
- (39) Lawson, C. L.; Hanson, R. J.; Kincaid, D. R.; Krogh, F. T. *ACM Trans. Math. Software* **1979**, *5*, 308–323.
- (40) Li, Y. S.; Wrinn, M. C.; Newsam, J. M.; Sears, M. P. *J. Comput. Chem.* **1995**, *16*, 226–234.
- (41) Baker, J.; Shirel, M. *Parallel Comput.* **2000**, *26*, 1011–1024.
- (42) Brown, S. T.; Kong, J. *Chem. Phys. Lett.* **2005**, *408*, 395–402.
- (43) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. *J. Comput. Phys.* **2009**, *228*, 8367–8379.
- (44) Izmaylov, A. F.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 104103.
- (45) White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1994**, *101*, 6593–6605.
- (46) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1994**, *230*, 8–16.
- (47) TURBOMOLE, developer version; a development of the University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007. <http://www.turbomole.com> (accessed Jul 25, 2011).
- (48) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (49) Pople, J. A.; Gill, P. M. W.; Johnson, B. G. *Chem. Phys. Lett.* **1992**, *199*, 557–560.
- (50) Towler, M. D.; Zupan, A.; Causà, M. *Comput. Phys. Commun.* **1996**, *98*, 181–205.
- (51) Kudin, K. N.; Scuseria, G. E. *Phys. Rev. B* **2000**, *61*, 16440–16453.
- (52) Pisani, C.; Dovesi, R.; Roetti, C. In *Hartree-Fock Ab Initio Treatment of Crystalline Systems*; Berthier, G., Dewar, M. J. S., Fischer, H., Fukui, K., Hall, G. G., Hinze, J., Jaffé, H. H., Jortner, J., Kutzelnigg, W., Ruedenberg, K., Tomasi, J., Eds.; Springer-Verlag: Berlin, 1988; Vol. 48, p 25.
- (53) Sierka, M.; Hogekamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136–9148.
- (54) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (55) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (56) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753–12762.

Quantum Electrodynamics Effects in Rovibrational Spectra of Molecular Hydrogen

Jacek Komasa,^{*,†} Konrad Piszczatowski,[‡] Grzegorz Łach,[‡] Michał Przybytek,[‡] Bogumił Jeziorski,[‡] and Krzysztof Pachucki[§]

[†]Faculty of Chemistry, A. Mickiewicz University, Grunwaldzka 6, 60-780 Poznań, Poland

[‡]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

[§]Faculty of Physics, University of Warsaw, Hoża 69, 00-681 Warsaw, Poland

 Supporting Information

ABSTRACT: The dissociation energies from all rovibrational levels of H₂ and D₂ in the ground electronic state are calculated with high accuracy by including relativistic and quantum electrodynamics (QED) effects in the nonadiabatic treatment of the nuclear motion. For D₂, the obtained energies have theoretical uncertainties of 0.001 cm⁻¹. For H₂, similar uncertainties are for the lowest levels, while for the higher ones the uncertainty increases to 0.005 cm⁻¹. Very good agreement with recent high-resolution measurements of the rotational $\nu = 0$ levels of H₂, including states with large angular momentum J , is achieved. This agreement would not have been possible without accurate evaluation of the relativistic and QED contributions and may be viewed as the first observation of the QED effects, mainly the electron self-energy, in a molecular spectrum. For several electric quadrupole transitions, we still observe certain disagreement with experimental results, which remains to be explained.

1. INTRODUCTION

Due to its simplicity, the hydrogen molecule and its isotopomers D₂ and HD are natural benchmark systems for testing various computational and experimental approaches in molecular spectroscopy. The measurements of their rovibrational levels have reached such a precision level^{1–4} that not only the nonadiabatic and the leading relativistic effects but also quantum electrodynamics (QED) and possibly finite nuclear size corrections become important. In this work, we exploit recent theoretical progress in calculating the spectra of H₂ and D₂ using explicitly correlated basis sets built of Gaussian or exponential functions. For the latter, exact analytic formulas have recently been derived,⁵ making possible the calculation of the Born–Oppenheimer potential with an uncertainty smaller than 10⁻⁹ cm⁻¹. The finite nuclear mass effects are calculated using nonadiabatic perturbation theory (NAPT).^{6,7} In this approach, leading nonadiabatic corrections are obtained by solving a suitable modification of the radial Schrödinger equation for the nuclear motion. The advantage of this procedure is that three universal nonadiabatic potentials entering this equation are sufficient to obtain all rovibrational levels supported by a given electronic state. At the precision level adequate for the present purposes, the relativistic, QED, and finite nuclear size effects can be taken into account at the adiabatic level of theory and included by means of appropriate corrections to the Born–Oppenheimer potential. These corrections are expectation values of the Breit–Pauli Hamiltonian,⁸ or effective QED operators,⁹ computed with the clamped nuclei nonrelativistic electronic wave function. Our approach is based on the expansion of energy levels in powers of fine structure constant α ,^{10,11} which was originally developed for hydrogenic systems and applied in highly accurate calculations for light atoms.^{12–15}

2. HISTORICAL PERSPECTIVE

A meaningful confrontation of theoretically predicted rovibrational levels with high-resolution spectroscopy measurements became possible in 1983 with the appearance of the pioneering work of Wolniewicz,¹⁶ who accurately computed the nonadiabatic corrections to all vibrational levels of H₂, HD, and D₂. For the H₂ molecule, Wolniewicz observed a significant disagreement of up to 0.7 cm⁻¹ with the experimental vibrational excitation energies available at that time.¹⁷ In a subsequent development, Kolos et al.¹⁸ recomputed the Born–Oppenheimer potential for H₂, obtaining the adiabatic levels 0.1–0.2 cm⁻¹ below the Wolniewicz values. Using the nonadiabatic corrections from the Wolniewicz work, the authors of ref 18 obtained improved agreement (with errors up to 0.2–0.3 cm⁻¹) when compared with a more recent measurement of Dabrowski.¹⁹

The works of Wolniewicz¹⁶ and of Kolos et al.¹⁸ were based on the relativistic corrections computed in 1964²⁰ using a rather small basis (of James and Coolidge type²¹) and only for internuclear distances $R \leq 3.7$ bohr. To eliminate this source of uncertainty, Wolniewicz²² recomputed in 1993 the relativistic corrections for a wide range of distances employing a large, asymptotically correct basis set introduced in 1966 by Kolos and Wolniewicz.²³ To obtain the nonadiabatic corrections, Wolniewicz used an *ab initio* based scaling and extrapolation procedure proposed by Schwartz and LeRoy,²⁴ which he believed to be more reliable (except for $\nu = J = 0$) than his *ab initio* data of 1983.¹⁶ This work led to some improvement in agreement with the experimental vibrational spectrum

Received: June 24, 2011

Published: August 25, 2011

of Dabrowski,¹⁹ as compared with the 1986 work of Kolos et al.¹⁸

As the method of Schwartz and LeRoy²⁴ had not been performing very satisfactorily in some applications, especially in the cases of HD and HT,²⁵ in 1995 Wolniewicz²⁶ recomputed the nonadiabatic corrections obtaining significantly more accurate values for higher vibrational levels. This resulted in further improvement of the agreement with the vibrational quanta of ref 19. It is difficult to say how significant this improvement was, since the disagreement with the experimental results of Dabrowski¹⁹ was on the last digit reported by her and the corresponding experimental uncertainty is not clear.

According to Wolniewicz,²⁶ the main source of uncertainty of his results is the QED corrections, which he evaluated only approximately. These corrections are accurately taken into account in the present calculations, with all α^3 and the leading α^4 terms included (α is the fine structure constant). Results for the ground rovibrational state (i.e., for the dissociation energy) of H₂ and D₂ have already been presented by us in ref 27 and for HD in ref 28. Very recently, Salumbides et al.²⁹ reported highly accurate experimental values of the rotational energies of $v = 0$ levels of H₂. The uncertainty of the majority of these levels ranges from 0.0001 cm⁻¹ up to 0.005 cm⁻¹. Such an accuracy, in connection with our calculations, is sufficient to determine the magnitude of QED effects directly from the measurements. It should be noted that QED corrections have also been treated for more complex systems, such as the water molecule,³⁰ but using an approximate one-electron approach; thus not all of the α^3 terms were taken into account. However, for such systems, the nonrelativistic treatment of the spectrum is still not accurate enough to see QED effects in a head-on comparison between theory and experiment.³¹

3. THE NONADIABATIC SCHRÖDINGER EQUATION

The nonrelativistic Schrödinger equation is solved using the NAPT approach.^{6,7} For electronic states of Σ symmetry, the zeroth-order approximation to the total wave function can be assumed to be a product of the electronic ϕ_{el} and nuclear χ functions (atomic units will be used throughout):

$$\phi_a(\vec{r}, \vec{R}) = \phi_{\text{el}}(\vec{r}; \vec{R}) \chi(\vec{R}) \quad (1)$$

The electronic wave function ϕ_{el} satisfies the clamped nuclei Schrödinger equation:

$$[H_{\text{el}} - \mathcal{E}_{\text{el}}(R)]\phi_{\text{el}} = 0 \quad (2)$$

where H_{el} is the electronic Hamiltonian, and thus parametrically depends on the internuclear distance R . The function χ satisfies the nuclear Schrödinger equation

$$\left[-\frac{\nabla_{\vec{R}}^2}{2\mu_{\text{n}}} + \mathcal{E}_{\text{el}}(R) + \mathcal{E}_{\text{a}}(R) \right] \chi(\vec{R}) = E_a \chi(\vec{R}) \quad (3)$$

with μ_{n} being the nuclear reduced mass, and the adiabatic correction $\mathcal{E}_{\text{a}}(R)$ is given by the electronic matrix element

$$\mathcal{E}_{\text{a}}(R) = \langle \phi_{\text{el}} | H_{\text{n}} | \phi_{\text{el}} \rangle_{\text{el}} \quad (4)$$

of the nuclear part $H_{\text{n}} = H - H_{\text{el}}$ of the total Hamiltonian H (the brackets $\langle \dots \rangle_{\text{el}}$ denote integration over the electronic coordinates only). For specified angular momentum J , the nuclear

equation becomes

$$\left[-\frac{1}{R^2} \frac{\partial}{\partial R} R^2 \frac{\partial}{\partial R} + \frac{J(J+1)}{2\mu_{\text{n}} R^2} + \mathcal{E}_{\text{a}}(R) + \mathcal{E}_{\text{el}}(R) \right] \chi_J(R) = E_a \chi_J(R) \quad (5)$$

Using NAPT,^{6,7} all of the finite nuclear mass corrections are obtained perturbatively in progressive powers of the electron–nuclear mass ratio. Up to the order $\mathcal{O}(\mu_{\text{n}}^{-2})$, they can all be included in the following radial equation as an R -dependent modification of the effective nuclear reduced mass and of the interaction potential:

$$\left[-\frac{1}{R^2} \frac{\partial}{\partial R} R^2 \frac{\partial}{\partial R} + \frac{J(J+1)}{2\mu_{\perp}(R) R^2} + \mathcal{Y}(R) \right] \tilde{\chi}_J(R) = E \tilde{\chi}_J(R) \quad (6)$$

where

$$\mathcal{Y}(R) = \mathcal{E}_{\text{el}}(R) + \mathcal{E}_{\text{a}}(R) + \delta \mathcal{E}_{\text{na}}(R) \quad (7)$$

The modifications of the effective nuclear reduced mass are of the form

$$\frac{1}{2\mu_{\parallel}(R)} \equiv \frac{1}{2\mu_{\text{n}}} + \frac{1}{\mu_{\text{n}}^2} \left\langle \vec{n} \cdot \vec{\nabla}_R \phi_{\text{el}} \left| \frac{1}{(\mathcal{E}_{\text{el}} - H_{\text{el}})'} \right| \vec{n} \cdot \vec{\nabla}_R \phi_{\text{el}} \right\rangle_{\text{el}} \quad (8)$$

and

$$\frac{1}{2\mu_{\perp}(R)} \equiv \frac{1}{2\mu_{\text{n}}} + \frac{1}{\mu_{\text{n}}^2} \sum_{ij} \frac{(\delta^{ij} - n^i n^j)}{2} \left\langle \nabla_R^i \phi_{\text{el}} \left| \frac{1}{\mathcal{E}_{\text{el}} - H_{\text{el}}} \right| \nabla_R^j \phi_{\text{el}} \right\rangle_{\text{el}} \quad (9)$$

with $\vec{n} = \vec{R}/R$ and the *prime* in the resolvent indicating the orthogonalization to ϕ_{el} . The concept of the variable nuclear reduced mass has been present in the literature for a long time now,^{32–35} and NAPT supplies explicit formulas for their numerical evaluation. The nonadiabatic correction $\delta \mathcal{E}_{\text{na}}(R)$ to the interaction potential can also be expressed in terms of the second-order electronic matrix elements. Since the corresponding formula is rather complicated, we refer the reader to refs 6 and 7 for its detailed form. The main advantage of the presented approach is that a single radial equation (eq 6) gives quite accurately all (nonrelativistic) rovibrational states of a diatomic molecule. In the present implementation of NAPT, the computational precision is limited by neglected $\mathcal{O}(\mu_{\text{n}}^{-5/2})$ corrections, which for the dissociation energy of the lowest rovibrational state are estimated to amount about 0.0001 cm⁻¹. Other important corrections are considered in the following section.

4. RELATIVISTIC AND QED CORRECTIONS

When computed using the adiabatic wave function of eq 1, the leading-order relativistic, QED, and finite nuclear size corrections enter only through an effective potential $\mathcal{V}(R)$, namely

$$\begin{aligned} \mathcal{V}(R) = & \mathcal{E}_{\text{el}}(R) + \mathcal{E}_a(R) + \delta\mathcal{E}_{\text{na}}(R) + \mathcal{E}^{(2)}(R) \\ & + \mathcal{E}^{(3)}(R) + \mathcal{E}^{(4)}(R) + \mathcal{E}_{\text{fs}}(R) \end{aligned} \quad (10)$$

The relativistic correction $\mathcal{E}^{(2)}(R)$ is the electronic expectation value of the Breit–Pauli Hamiltonian,⁸ which for a Σ state of a many-electron molecule is

$$\begin{aligned} \mathcal{E}^{(2)}(R) = & \alpha^2 \left\langle \phi_{\text{el}} \left| -\frac{1}{8} \sum_a p_a^4 + \frac{\pi}{2} \sum_{a,A} Z_A \delta(\vec{r}_{aA}) + \pi \sum_{a < b} \delta(\vec{r}_{ab}) \right. \right. \\ & \left. \left. - \frac{1}{2} \sum_{a < b} \left(\vec{p}_a \frac{1}{r_{ab}} \vec{p}_b + \vec{p}_a \cdot \vec{r}_{ab} \frac{1}{r_{ab}^3} \vec{r}_{ab} \cdot \vec{p}_b \right) \right| \phi_{\text{el}} \right\rangle \end{aligned} \quad (11)$$

where the indices a and A correspond to electrons and nuclei, respectively.

The leading QED correction for a Σ state is

$$\begin{aligned} \mathcal{E}^{(3)}(R) = & \alpha^3 \sum_{a < b} \left\{ \left[\frac{164}{15} + \frac{14}{3} \ln \alpha \right] \langle \phi_{\text{el}} | \delta(\vec{r}_{ab}) | \phi_{\text{el}} \rangle_{\text{el}} \right. \\ & \left. - \frac{7}{6\pi} \left\langle \phi_{\text{el}} \left| \frac{1}{r_{ab}^3} \right| \phi_{\text{el}} \right\rangle_{\text{el}} \right\} \\ & + \alpha^3 \sum_{a,A} \left[\frac{19}{30} - 2 \ln \alpha - \ln k_0(R) \right] \frac{4Z_A}{3} \langle \phi_{\text{el}} | \delta(\vec{r}_{aA}) | \phi_{\text{el}} \rangle_{\text{el}} \end{aligned} \quad (12)$$

The matrix element of $1/r_{ab}^3$ requires subtraction of a divergence at $r_{ab} = 0$; for details, see ref 9. In the adiabatic approximation, the R -dependent Bethe logarithm $\ln k_0(R)$ is defined by²⁷

$$\ln k_0(R) = \frac{\langle \phi_{\text{el}} | \sum_a \vec{p}_a (H_{\text{el}} - \mathcal{E}_{\text{el}}) \ln[2(H_{\text{el}} - \mathcal{E}_{\text{el}})] \sum_b \vec{p}_b | \phi_{\text{el}} \rangle}{\langle \phi_{\text{el}} | \sum_a \vec{p}_a (H_{\text{el}} - E_{\text{cl}}) \sum_b \vec{p}_b | \phi_{\text{el}} \rangle} \quad (13)$$

The α^4 and higher-order QED corrections^{36,37} are in general not known for molecules. One can however approximate them by the numerically dominating one-loop self-energy and vacuum polarization corrections known from the hydrogenic Lamb shift:³⁶

$$\mathcal{E}_{\text{one-loop}}^{(4)}(R) = \pi\alpha^4 \left(\frac{427}{96} - \ln 4 \right) \sum_{a,A} Z_A \langle \phi_{\text{el}} | \delta(\vec{r}_{aA}) | \phi_{\text{el}} \rangle_{\text{el}} \quad (14)$$

On the basis of the atomic calculations,^{36,37} we conservatively estimate that this formula approximates the accurate value of $\mathcal{E}^{(4)}$ with an error of at most 50%.

The finite nuclear size correction $\mathcal{E}_{\text{fs}}(R)$ is given, to a very good approximation, by the root-mean-square nuclear charge radius r_{ch} :

$$\mathcal{E}_{\text{fs}}(R) = \frac{2\pi}{3} \alpha^2 \sum_{a,A} Z_A \frac{r_{\text{ch}}^2(A)}{\lambda_C^2} \langle \phi_{\text{el}} | \delta(\vec{r}_{aA}) | \phi_{\text{el}} \rangle_{\text{el}} \quad (15)$$

where $\lambda_C = 386.15926459$ fm is the Compton wavelength over 2π ($r_{\text{ch}}(\text{H}) = 0.84184(67)$ fm and $r_{\text{ch}}(\text{D}) = 2.1402(28)$ fm). For H_2 , the energy level shift resulting from this correction is always less than 0.0001 cm^{-1} , whereas for D_2 it amounts to only 0.0002 cm^{-1} or less. Nevertheless, in both cases, this correction has been included in our final results. According to our knowledge, there are no further corrections to the rovibrational energies which contribute above 0.001 cm^{-1} . The atomic hyperfine splittings are larger than that but are subtracted from the experimental dissociation energies. Recently reported³⁸ *gerade*–*ungerade* mixing and splitting effects in H_2 turned out to be smaller than 10^{-6} cm^{-1} and thus are entirely negligible for the present purposes.

5. NUMERICAL APPROACH

A very accurate clamped nuclei potential for the $X^1\Sigma_g^+$ state was reported recently in ref 5. For the whole energy curve, an accuracy on the order of 10^{-9} cm^{-1} has been reached. This is the most accurate potential to date for H_2 itself but also for any molecular system with two or more electrons. Increasing the accuracy to this level has been possible thanks to the discovery of analytic formulas for two-center two-electron integrals with exponential functions.³⁹ To achieve this high numerical accuracy, different basis sets were used, depending on the internuclear distance R . For $R < 12$ bohr, the James–Coolidge basis functions²¹ of the form

$$\begin{aligned} \psi_k(\vec{r}_1, \vec{r}_2) = & (1 + \hat{P}_{12})(1 + \hat{i}) \\ & \exp(-\alpha(r_{1A} + r_{1B}) - \alpha(r_{2A} + r_{2B})) \\ & \times r_{12}^{n_{1k}} (r_{1A} - r_{1B})^{n_{2k}} (r_{2A} - r_{2B})^{n_{3k}} (r_{1A} + r_{1B})^{n_{4k}} (r_{2A} + r_{2B})^{n_{5k}} \end{aligned} \quad (16)$$

have been employed. The symmetry projector $(1 + \hat{P}_{12})$ ensures a singlet state, while the spatial projector $(1 + \hat{i})$ ensures the *gerade* symmetry. Since in the actual numerical calculations one can use only a finite number of basis functions, one has to somehow select the most appropriate finite subset of functions of eq 16. We assumed, therefore, that the finite basis consists of all functions ψ_k with nonnegative integers n_{ik} such that

$$\sum_{i=1}^5 n_{ik} \leq \Omega \quad (17)$$

with $\Omega = 3-20$, and the final result is obtained by a numerical extrapolation with $\Omega \rightarrow \infty$. The nonlinear parameters were optimized separately for each internuclear distance R , and then the exponential convergence to a complete basis set as $\Omega \rightarrow \infty$ was observed.

To represent the electronic wave function at $12 \leq R \leq 20$ bohr, a special case of the Kolos and Wolniewicz basis⁴⁰ was used

$$\begin{aligned} \psi_k(\vec{r}_1, \vec{r}_2) = & (1 + \hat{P}_{12})(1 + \hat{i}) \exp(-(r_{1A} + r_{2B})) r_{12}^{n_{1k}} r_{1A}^{n_{2k}} r_{1B}^{n_{3k}} r_{2A}^{n_{4k}} r_{2B}^{n_{5k}} \end{aligned} \quad (18)$$

Table 1. Theoretically Predicted Dissociation Energies (in cm^{-1}) of All 302 Bound States of H_2^a

$\nu \setminus J$	0	1	2	3	4	5	6	7
0	36118.0696	35999.5827	35763.6964	35412.5507	34949.2721	34377.8800	33703.1726	32930.5991
1	31956.9034	31844.3290	31620.2311	31286.6781	30846.6902	30304.1479	29663.6809	28930.5440
2	28031.0670	27924.2753	27711.7087	27395.3676	26978.1720	26463.8724	25856.9411	25162.4519
3	24335.6787	24234.5818	24033.3726	23733.9878	23339.2543	22852.8015	22278.9563	21622.6265
4	20867.7039	20772.2598	20582.3263	20299.7818	19927.3669	19468.6008	18927.6785	18309.3586
5	17626.1400	17536.3593	17357.7260	17092.0653	16742.0411	16311.0741	15803.2433	15223.1787
6	14612.2901	14528.2461	14361.0627	14112.5194	13785.2138	13382.4836	12908.3120	12367.2247
7	11830.1543	11751.9964	11596.5661	11365.6062	11061.6626	10688.0095	10248.5600	9747.7694
8	9286.9790	9214.9526	9071.7726	8859.1575	8579.6206	8236.3999	7833.3746	7374.9763
9	6994.0292	6928.5057	6798.3271	6605.2057	6351.6504	6040.9022	5676.8602	5264.0027
10	4967.6786	4909.1996	4793.1191	4621.1702	4395.9005	4120.6166	3799.3225	3436.6592
11	3230.9712	3180.3202	3079.9265	2931.5894	2737.9677	2502.5403	2229.5680	1924.0695
12	1815.8955	1774.2213	1691.8517	1570.7364	1413.7847	1224.8610	1008.8062	771.5140
13	766.7551	735.8177	675.0810	586.8399	474.5979	343.1936	199.1012	51.2532
14	144.7964	127.6357	94.9453	50.2393	0.0265 ^b			
$\nu \setminus J$	8	9	10	11	12	13	14	15
0	32066.1266	31116.1067	30087.1493	28986.0061	27819.4682	26594.2780	25317.0575	23994.2505
1	28110.4874	27209.6288	26234.3321	25191.0967	24086.4605	22926.9171	21718.8492	20468.4759
2	24385.9548	23533.3534	22610.7880	21624.5312	20580.8960	19486.1590	18346.4985	17167.9474
3	20889.1803	20084.3279	19214.0114	18284.3049	17301.3289	16271.1794	15199.8715	14093.2980
4	17618.8479	16861.6882	16043.6518	15170.6479	14248.6434	13283.5982	12281.4160	11247.9099
5	14575.9499	13866.9601	13101.8471	12286.3973	11426.4734	10527.9578	9596.7126	8638.5554
6	11764.1860	11104.4981	10393.7108	9637.5433	8841.8215	8012.4332	7155.3007	6276.3735
7	9190.5379	8582.1192	7928.0390	7234.0292	6505.9792	5749.9080	4971.9605	4178.4321
8	6866.1012	6312.0299	5718.3612	5090.9637	4435.9509	3759.6871	3068.8309	2370.4348
9	4807.3140	4312.2229	3784.5601	3230.5444	2656.8077	2070.4770	1479.3467	892.2132
10	3037.8559	2608.7032	2155.5628	1685.4344	1206.1180	726.5520	257.5346	
11	1591.8264	1239.4455	874.5218	506.0045	145.0445			
12	520.1248	263.4848	13.3741					
$\nu \setminus J$	16	17	18	19	20	21	22	23
0	22632.0799	21236.5171	19813.2627	18367.7376	16905.0813	15430.1579	13947.5677	12461.6638
1	19181.8142	17864.6524	16522.5351	15160.7573	13784.3676	12398.1776	11006.7791	9614.5672
2	15956.3587	14717.3851	13456.4677	12178.8362	10889.5174	9593.3530	8295.0256	6999.0945
3	12957.2014	11797.1591	10618.5796	9426.7112	8226.6608	7023.4255	5821.9382	4627.1316
4	10188.7828	9109.6212	8015.9040	6913.0252	5806.3338	4701.1940	3603.0755	2517.6864
5	7659.2518	6664.5260	5660.0900	4651.6971	3645.2264	2646.8145	1663.0640	701.3929
6	5381.6434	4477.1860	3569.2356	2664.3109	1769.4191	892.4054	42.5995	
7	3375.8309	2570.9911	1771.2705	984.8942	221.6033			
8	1672.1303	982.4662	311.5626					
9	319.5464							
$\nu \setminus J$	24	25	26	27	28	29	30	31
0	10976.5735	9496.2237	8024.3709	6564.6364	5120.5483	3695.5928	2293.2796	917.2285
1	8225.7696	6844.4839	5474.7238	4120.4784	2785.7900	1474.8629	192.2238	
2	5710.0438	4432.3453	3170.5449	1929.3851	713.9904			
3	3444.0280	2277.8701	1134.3209	19.7931 ^b				
4	1451.1807	410.5020						

^a ν and J are the vibrational and rotational quantum numbers, respectively. Error estimation for individual states is given in the Supporting Information.

^b This state is entirely due to nonadiabatic effects. Since it does not appear at the adiabatic approximation, its energy may be substantially less accurate than the energies of the remaining rovibrational states.

with Ω up to 16. This basis, which may be viewed as a generalized Heitler–London basis (it contains the Heitler–London function),

is needed to correctly describe molecular dissociation. At $R = 12$ bohr, the accuracy achieved with the basis set of eq 18 is close to

Table 2. Theoretically Predicted Dissociation Energies (in cm^{-1}) of All 598 Bound States of D_2^a

$\nu \setminus J$	0	1	2	3	4	5	6	7	8	9	10
0	36748.3634	36688.5828	36569.2963	36391.0490	36154.6478	35861.1488	35511.8412	35108.2289	34652.0088	34145.0488	33589.3635
1	33754.7463	33697.0758	33582.0029	33410.0597	33182.0337	32898.9554	32562.0824	32172.8812	31733.0063	31244.2778	30708.6583
2	30880.2430	30824.6456	30713.7129	30547.9643	30328.1686	30055.3317	29730.6810	29355.6472	28931.8440	28461.0461	27945.1666
3	28122.7615	28069.2076	27962.3556	27802.7132	27591.0311	27328.2917	27015.6934	26654.6328	26246.6848	25793.5810	25297.1880
4	25480.6486	25429.1157	25326.3000	25172.6976	24969.0422	24716.2938	24415.6231	24068.3944	23676.1459	23240.5687	22763.4852
5	22952.7114	22903.1852	22804.3772	22656.7728	22461.0894	22218.2657	21929.4464	21595.9654	21219.3265	20801.1827	20343.3156
6	20538.2479	20490.7226	20395.9113	20254.2889	20066.5579	19833.6368	19556.6462	19236.8915	18875.8444	18475.1232	18036.4718
7	18237.0868	18191.5664	18100.7600	17965.1331	17785.3739	17562.3826	17297.2573	16991.2773	16645.8854	16262.6681	15843.3359
8	16049.6405	16006.1398	15919.3685	15789.7833	15618.0596	15405.0808	15151.9249	14859.8479	14530.2666	14164.7400	13764.9495
9	13976.9722	13935.5186	13852.8377	13729.3782	13565.8044	13362.9849	13121.9801	12844.0261	12530.5179	12182.9911	11803.1038
10	12020.8819	11981.5173	11903.0111	11785.8056	11630.5553	11438.1169	11209.5363	10946.0334	10648.9863	10319.9134	9960.4560
11	10184.0153	10146.7989	10072.5866	9961.8154	9815.1328	9633.3864	9417.6115	9169.0172	8888.9705	8578.9799	8240.6789
12	8470.0026	8435.0143	8365.2566	8261.1634	8123.3777	7952.7421	7750.2867	7517.2160	7254.8939	6964.8287	6648.6575
13	6883.6348	6850.9801	6785.8889	6688.7948	6560.3401	6401.3676	6212.9089	5996.1728	5752.5308	5483.5039	5190.7489
14	5431.0906	5400.9066	5340.7582	5251.0815	5132.5237	4985.9347	4812.3575	4613.0166	4389.3065	4142.7793	3875.1345
15	4120.2289	4092.6936	4037.8462	3956.1303	3848.2050	3714.9379	3557.3966	3376.8392	3174.7048	2952.6049	2712.3169
16	2960.9710	2936.3160	2887.2361	2814.1894	2717.8581	2599.1427	2459.1559	2299.2164	2120.8431	1925.7521	1715.8578
17	1965.8035	1944.3335	1901.6362	1838.1950	1754.7324	1652.2069	1531.8105	1394.9679	1243.3394	1078.8290	903.6027
18	1150.4486	1132.5743	1097.0897	1044.5231	975.6684	891.5869	793.6143	683.3726	562.7936	434.1618	300.1927
19	534.7343	521.0330	493.9347	454.0526	402.3180	339.9993	268.7360	190.6028	108.2387	25.1501	
20	143.4642	134.8278	117.9548	93.6691	63.2806	28.7401					
21	1.6642	0.0491 ^b									
$\nu \setminus J$	11	12	13	14	15	16	17	18	19	20	21
0	32987.0905	32340.4669	31651.8060	30923.4758	30157.8783	29357.4315	28524.5520	27661.6407	26771.0694	25855.1707	24916.2279
1	30128.2296	29505.1695	28841.7298	28140.2145	27402.9610	26632.3216	25830.6476	25000.2755	24143.5140	23262.6347	22359.8626
2	27386.2346	26786.3718	26147.7718	25472.6790	24763.3702	24022.1367	23251.2697	22453.0464	21629.7184	20783.5020	19916.5706
3	24759.4845	24182.5401	23568.4939	22919.5347	22237.8832	21525.7751	20785.4464	20019.1210	19228.9994	18417.2497	17586.0005
4	22246.8277	21692.6167	21102.9411	20479.9386	19825.7789	19142.6472	18432.7304	17698.2055	16941.2288	16163.9280	15368.3950
5	19847.6143	19316.0545	18750.6790	18153.5798	17526.8807	16872.7226	16193.2502	15490.6008	14766.8946	14024.2276	13264.6654
6	17561.7401	17052.8634	16511.8438	15940.7327	15341.6146	14716.5931	14067.7784	13397.2768	12707.1824	11999.5701	11276.4914
7	15389.7037	14903.6716	14387.2068	13842.3268	13271.0842	12675.5536	12057.8202	11419.9700	10764.0828	10092.2265	9406.4537
8	13332.6805	12869.8040	12378.2591	11860.0373	11317.1686	10751.7089	10165.7299	9561.3111	8940.5334	8305.4757	7658.2131
9	11392.6185	10953.3842	10487.3206	9996.4033	9482.6509	8948.1137	8394.8655	7824.9967	7240.6106	6643.8215	6036.7566
10	9572.3616	9157.4667	8717.6828	8254.9820	7771.3865	7268.9587	6749.7950	6216.0214	5669.7926	5113.2944	4548.7496
11	7875.8101	7486.2101	7073.7960	6640.5542	6188.5304	5719.8236	5236.5818	4741.0024	4235.3355	3721.8929	3203.0627
12	6308.1321	5945.1055	5561.5208	5159.4021	4740.8480	4308.0294	3863.1902	3408.6539	2946.8363	2480.2668	2011.6212
13	4876.0459	4541.2881	4188.4733	3819.6987	3437.1601	3043.1557	2640.0964	2230.5252	1817.1485	1402.8863	990.9510
14	3588.2093	3283.9721	2964.5198	2632.0792	2289.0159	1937.8510	1581.2918	1222.2810	864.0776	510.3933	165.6395
15	2455.7797	2185.0946	1902.5323	1610.5491	1311.8159	1009.2686	706.1933	406.3736	114.3694		
16	1493.2790	1260.3555	1019.6781	774.1419	527.0394	282.2304	44.4834				
17	720.1223	531.2062	340.1405	150.8950							
18	164.1866	30.3695									
$\nu \setminus J$	22	23	24	25	26	27	28	29	30	31	32
0	23956.4683	22978.0571	21983.0935	20973.6071	19951.5567	18918.8293	17877.2405	16828.5352	15774.3900	14716.4151	13656.1575
1	21437.3699	20497.2704	19541.6158	18572.3932	17591.5237	16600.8620	15602.1976	14597.2554	13587.6991	12575.1334	11561.1078
2	19031.0480	18129.0038	17212.4502	16283.3399	15343.5655	14394.9594	13439.2953	12478.2906	11513.6089	10546.8645	9579.6270
3	16737.3351	15873.2873	14995.8389	14106.9182	13208.4000	12302.1065	11389.8096	10473.2340	9554.0614	8633.9359	7714.4708
4	14556.6818	13730.7962	12892.7006	12044.3114	11187.4992	10324.0917	9455.8769	8584.6075	7712.0074	6839.7795	5969.6154
5	12490.2391	11702.9433	10904.7348	10097.5334	9283.2241	8463.6608	7640.6720	6816.0677	5991.6492	5169.2212	4350.6073
6	10539.9706	9792.0047	9034.5628	8269.5895	7499.0087	6724.7301	5948.6581	5172.7034	4398.7986	3628.9187	2865.1080
7	8708.8007	8001.2872	7285.9202	6564.6983	5839.6198	5112.6937	4385.9547	3661.4828	2941.4313	2228.0638	1523.8081
8	7000.8179	6335.3630	5663.9281	4988.6099	4311.5357	3634.8825	2960.9045	2291.9703	1630.6160	979.6255	342.1539

Table 2. Continued

$\nu \setminus J$	22	23	24	25	26	27	28	29	30	31	32
9	5421.5600	4800.4013	4175.4879	3549.0832	2923.5329	2301.3016	1685.0260	1077.5947	482.2723		
10	3978.4290	3404.6679	2829.8912	2256.6490	1687.6692	1125.9370	574.8194	38.2756			
11	2681.3328	2159.3251	1639.8473	1125.9722	621.1627	129.4853					
12	1543.7728	1079.8706	623.4674	178.7378							
13	584.9779	189.2547									
$\nu \setminus J$	33	34	35	36	37	38	39	40	41	42	43
0	12595.1047	11534.6886	10476.2901	9421.2440	8370.8452	7326.3541	6289.0044	5260.0106	4240.5776	3231.9113	2235.2318
1	10547.1215	9534.6279	8525.0406	7519.7400	6520.0811	5527.4021	4543.0351	3568.3189	2604.6139	1653.3230	715.9165
2	8613.4269	7649.7624	6690.1076	5735.9218	4788.6608	3849.7912	2920.8076	2003.2555	1098.7610		
3	6797.2565	5883.8701	4975.8879	4074.9000	3182.5291	2300.4554	1430.4494	574.4182			
4	5103.2079	4242.2662	3388.5358	2543.8247	1710.0382	889.2289					
5	3537.6706	2732.3404	1936.6489	1152.7828	383.1585						
6	2109.5182	1364.4607	632.4843								
7	831.3351	153.6817									

^a ν and J are the vibrational and rotational quantum numbers, respectively. Error estimation for individual states is given in the Supporting Information. ^b This state does not exist in the Born–Oppenheimer approximation and its energy may be substantially less accurate than the energies of the remaining rovibrational states.

that obtained with the James–Coolidge basis. The Born–Oppenheimer energies obtained as described above were fitted with an analytic function of the form given in ref 27. We estimate that it represents the Born–Oppenheimer potential with accuracy better than 10^{-5} cm^{-1} . Fortran code to compute this analytic function can be supplied upon request.

For the evaluation of adiabatic, nonadiabatic, relativistic, and QED corrections, we have not used exponential functions, like Kolos and Wolniewicz in refs 40 and 22, because we have not yet developed integrals with inverse quadratic powers of electronic variables. Instead, we used explicitly correlated Gaussian (ECG) functions of the form

$$\begin{aligned} \psi_k(\vec{r}_1, \vec{r}_2) \\ = (1 + \hat{P}_{12})(1 \pm \hat{i}) \exp\left[-\sum_{i,j=1}^2 A_{k,ij}(\vec{r}_i - \vec{s}_{k,i})(\vec{r}_j - \vec{s}_{k,j})\right] \end{aligned} \quad (19)$$

where the symmetric 2×2 matrices \mathbf{A}_k and vectors \vec{s}_k (lying on the internuclear axis) contain nonlinear parameters, five per basis function, to be variationally optimized. The 1200-term ECG bases were optimized with respect to \mathcal{E}_{el} for R spread over the range 0–12 bohr and employed to evaluate all of the corrections. At larger values of R , the relativistic and QED corrections to the potential were represented using the asymptotic constants reported in refs 41 and 27. Details of numerical evaluations of adiabatic and nonadiabatic corrections were presented in refs 6 and 7, while the evaluation of relativistic and QED corrections to the potential was presented in ref 27.

It should be stressed that the method of computing the rovibrational levels employed in this work differs slightly from that of ref 27. In this reference, the relativistic and QED corrections were obtained directly by averaging the corrections to the potential with nuclear wave functions obtained in the adiabatic approximation. In the present work, the relativistic and QED corrections to the potential are used in the nonadiabatic equation for the nuclear motion, so we can deal with exceptional states which do not exist on the adiabatic level. The

corresponding corrections to the rovibrational levels were obtained as appropriate differences in calculated energies. Thus, the present “ α^2 ” energies contain very small contributions of the fourth and higher order in α as well as the first and higher order in α^2/μ_n . Similarly, the “ α^3 ” energies contain some very small contributions from the fifth and higher powers of α . For low values of the vibrational quantum number ν , both approaches give results differing by 0.0001 cm^{-1} or less. At the highest values of ν , apart from exceptional states described later on, the differences reach 0.0005 cm^{-1} , i.e., are of the order of the neglected recoil corrections. These differences appear only at the relativistic level. The QED corrections differ always less than 0.0001 cm^{-1} .

6. RESULTS FOR HYDROGEN AND DEUTERIUM MOLECULES

In Tables 1 and 2, we present the complete rovibrational spectrum of molecular hydrogen and deuterium in the electronic ground state ($X^1\Sigma_g^+$) in terms of dissociation energies with a precision of about 0.001 cm^{-1} for D_2 and from 0.001 to 0.005 cm^{-1} for H_2 . The individual adiabatic, nonadiabatic, relativistic, and QED contributions for each level of H_2 and D_2 are listed in the Supporting Information files associated with this article (the very small finite-size contribution is included in the relativistic correction). The estimated theoretical uncertainties corresponding to each level are also listed in these files. The largest uncertainty comes from the α^4 and higher-order QED corrections, which have been included only approximately, see eq 14, and from the higher-order $\mathcal{O}(\mu_n^{-5/2})$ NAPT corrections. Other significant sources of uncertainty are the neglected nonadiabatic relativistic effects. To estimate the uncertainty of each rovibrational level, we used the following procedure. We assumed that the error caused by the neglect of the relativistic recoil term can be estimated as m_e/μ_n times the α^2 correction and, analogously, times the α^3 correction to account for the missing QED recoil term. Similarly, we calculate the contribution to the error budget from the missing higher-order nonadiabatic terms as proportional to m_e/μ_n times the second-order

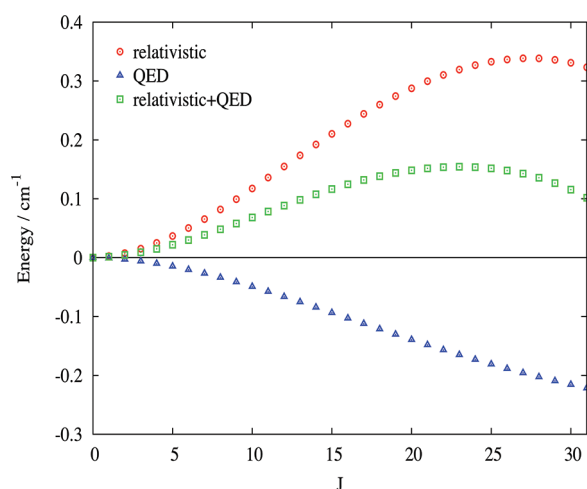


Figure 1. J dependence of the relativistic and QED contributions to the rotational excitation energies at $\nu = 0$. The relativistic and QED corrections to the ground, $J = 0$ level are $0.5318(5) \text{ cm}^{-1}$ and $0.1964(8) \text{ cm}^{-1}$, respectively.

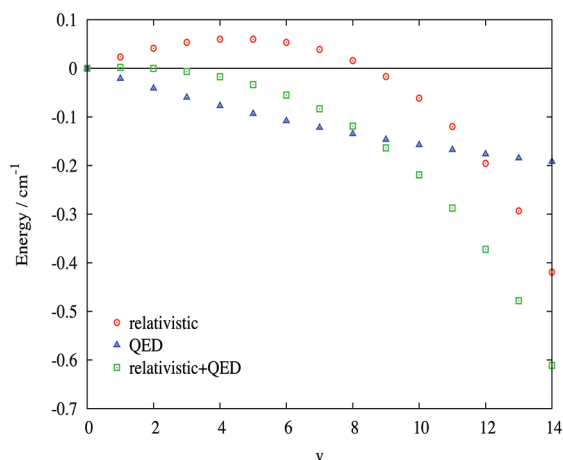


Figure 2. Vibrational quantum number dependence of the relativistic and QED contributions to the excitation energies at $J = 0$. The relativistic and QED corrections to the ground, $\nu = 0$ level are $0.5318(5) \text{ cm}^{-1}$ and $0.1964(8) \text{ cm}^{-1}$, respectively.

nonadiabatic correction. Another part of the uncertainty comes from the incomplete treatment of the higher-order QED effects. As previously,²⁷ we conservatively estimate that the $\mathcal{O}^{(4)}$ terms other than the one-loop term of eq 14 as well as all higher-order $\mathcal{O}^{(n)}$ terms, $n > 4$, contribute at most 50% of the value of $\mathcal{O}^{(4)}_{\text{one-loop}}$. The quadratic sum of the four error components described above leads to the overall uncertainty on D_0 or any other computed energy difference.

Among the levels listed in Tables 1 and 2, we note a couple of curiosities—the bound levels, which are absent in the Born–Oppenheimer approximation. Such levels do not reveal themselves until the finite mass corrections are taken into account. In H_2 , there are two such “nonadiabatic” levels: $(\nu, J) = (3, 27)$ and $(14, 4)$, with dissociation energies of about 19.79 cm^{-1} and 0.03 cm^{-1} , respectively. The former state was observed by Dabrowski¹⁹ as lying 36098.04 cm^{-1} above the $(0, 0)$ level, which corresponds to $D_0 = 20.03 \text{ cm}^{-1}$. The latter one, located $36\,118.12 \text{ cm}^{-1}$ above the reference state, she describes as “the

Table 3. Dissociation Energies (in cm^{-1}) of H_2 and D_2 in Their Ground State—A Comparison with the Experiment

	D_0/cm^{-1}	
	H_2	D_2
experiment (1993)	$36118.06(4)^a$	$36748.32(7)^a$
experiment (2004)	$36118.062(10)^b$	$36748.343(10)^b$
experiment (2009/10)	$36118.06962(37)^c$	$36748.36286(68)^d$
theory	$36118.0696(11)$	$36748.3634(9)$
difference	$0.0000(12)$	$0.0005(11)$

^a Ref 66. ^b Ref 67. ^c Ref 1. ^d Ref 2.

Table 4. The Lowest Rotational and Vibrational Excitation Energies (in cm^{-1})

	H_2	
	$J = 0 \rightarrow 1$	$\nu = 0 \rightarrow 1$
theory	$118.486812(9)$	$4161.1661(9)$
experiment	$118.48684(10)^a$	$4161.1660(3)^b$
difference	$-0.00003(10)$	$0.0001(9)$
	D_2	
	$J = 0 \rightarrow 1$	$\nu = 0 \rightarrow 1$
theory	$59.780615(3)$	$2993.6171(2)$
experiment	$59.78130(95)^c$	$2993.6130(19)^d$
difference	$-0.00068(95)$	$0.0041(19)$

^a Ref 68. ^b Ref 46. ^c Ref 2. ^d Ref 45.

last observed” and “presumably bound”. However, if we assign the up-to-date value of D_0 to the $(0, 0)$ level, the state observed by Dabrowski becomes a resonance located 0.05 cm^{-1} above the dissociation threshold. This value is an order of magnitude smaller than the accuracy declared by Dabrowski; therefore, from the experimental point of view, the question whether this state is bound or not remains open. In the D_2 spectrum, we predict the existence of only one such level $(21, 1)$ with $D_0 = 0.05 \text{ cm}^{-1}$. In this case, it is the adiabatic correction which makes this state bound. We note also the presence of another level located just below the dissociation threshold—the $(21, 0)$ state with $D_0 = 1.66 \text{ cm}^{-1}$. This state, in contrast, is a regular one, accommodated already by the Born–Oppenheimer potential.

The ν and J dependence of the relativistic and QED corrections to the rotational and vibrational excitation energies (relative the ground $\nu = 0, J = 0$ level) is shown in Figures 1 and 2. It may be pointed out that individual terms in eqs 11 and 12, like, e.g., mass velocity, Darwin, Breit or Araki-Sucher ones, lead to a monotonic dependence on ν or J . The nonmonotonic behavior observed in Figures 1 and 2 is due to cancellation effects and can be rationalized on the basis of the different R dependence of the corresponding corrections to the potential. One may observe that the QED correction is not much smaller than the relativistic one and that the neglect of the former would lead to a qualitatively incorrect ν or J dependence of the relativistic + QED contribution.

Theoretical predictions for the ground rovibrational state of H_2 and D_2 have already been presented by us in ref 27. Here, in Table 3, we compare these results with the most recent

Table 5. Energy of Rotational Excitations $\Delta E(J)$ in H_2 ($\nu = 0$) with Respect to the Ground Level ($\nu = 0, J = 0$)—A Comparison with the Experimental Results of Salumbides et al.^{27a}

J	$\Delta E_{\text{theo}}(J)$	$\Delta E_{\text{exptl}}(J)$	$\delta(\text{theo} - \text{exptl})$	$\Delta E_{\text{theo}}^{\text{QED}}(J)$	$\Delta E_{\text{exptl}}^{\text{QED}}(J)$
1	118.486812(9)	118.48684(10)	-0.00003(10)	-0.001030(4)	-0.00100(10)
2	354.37313(3)	354.3733(2)	-0.0002(2)	-0.00307(1)	-0.0029(2)
3	705.51883(5)	705.5189(3)	-0.0001(3)	-0.00610(2)	-0.0060(3)
4	1168.79743(9)	1168.7982(2)	-0.0008(2)	-0.01005(4)	-0.0093(2)
5	1740.1895(1)	1740.1895(3)	-0.0000(3)	-0.01487(6)	-0.0148(3)
6	2414.8970(2)	2414.898(5)	-0.001(5)	-0.02050(8)	-0.019(5)
7	3187.4705(2)	3187.472(5)	-0.002(5)	-0.0268(1)	-0.025(5)
8	4051.9430(3)	4051.943(5)	-0.000(5)	-0.0338(1)	-0.034(5)
9	5001.9628(4)	5001.963(5)	-0.000(5)	-0.0414(2)	-0.041(5)
10	6030.9202(5)	6030.921(5)	-0.001(5)	-0.0494(2)	-0.049(5)
11	7132.0634(6)	7132.066(5)	-0.003(5)	-0.0578(2)	-0.055(5)
12	8298.6014(6)	8298.600(5)	0.001(5)	-0.0665(3)	-0.068(5)
13	9523.7916(7)	9523.794(7)	-0.002(7)	-0.0754(3)	-0.073(7)
14	10801.0121(9)	10801.008(9)	0.004(9)	-0.0846(3)	-0.089(9)
15	12123.819(1)	12123.83(1)	-0.01(1)	-0.0938(4)	-0.08(1)
16	13485.990(1)	13485.99(1)	-0.00(1)	-0.1030(4)	-0.10(1)

^aThe two last columns show contributions to this energy from the QED effects: $\Delta E_{\text{theo}}^{\text{QED}}(J)$, predicted in our calculations, and $\Delta E_{\text{exptl}}^{\text{QED}}(J)$, extracted from the experimental data by subtracting our relativistic energies from the measured values given in column $\Delta E_{\text{exptl}}(J)$. All entries in cm^{-1} .

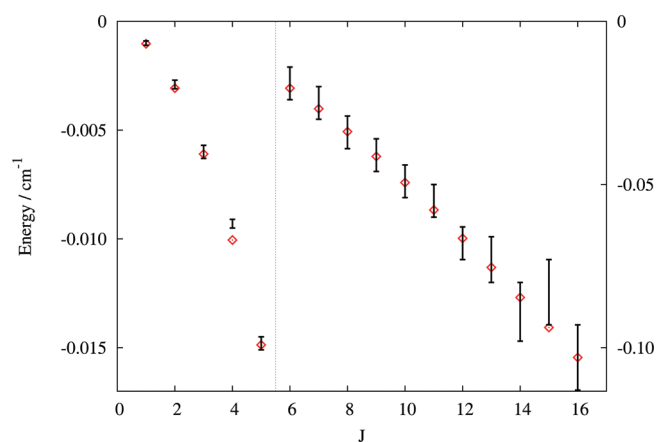


Figure 3. QED contributions to the rotational excitation energies at $\nu = 0$. Comparison of the present theoretical calculations (open diamonds) with the experimental data of ref 29. The left/right energy scale is for the levels to the left/right from the vertical line. The extent of the vertical bars shows the experimental uncertainties (listed in Table 5). The theoretical uncertainties (not shown) are at least an order of magnitude smaller than the experimental ones, except for $J = 4$ and $J = 5$, when they are 5 times smaller.

experimental data. Our present theoretical value of the dissociation energy of H_2 , $D_0 = 36\,118.0696(11) \text{ cm}^{-1}$, differing very slightly from that of ref 27, agrees very well with the value $36\,118.06962(37) \text{ cm}^{-1}$ derived experimentally by Liu et al.¹ The dissociation energies for D_2 are $36\,748.3634(9) \text{ cm}^{-1}$ from theory and $36\,748.36286(68) \text{ cm}^{-1}$ from the most recent experiment.² The perfect agreement with the sophisticated high-resolution measurements (see also ref 42 for a confirmation of the result from ref 2) indicates that we control numerical accuracy of the included energy contributions as well as all physical effects down to 0.001 cm^{-1} . It is worth emphasizing that this excellent agreement would have been impossible

Table 6. Vibrational Excitation Energies $E(\nu, J = 0) - E(\nu = 0, J = 0)$ for H_2 and Their Differences with the Experimental Values of Dabrowski¹⁹ (δ_{exptl}) and with the Best Previous Theoretical Values by Wolniewicz²⁶ (δ_{Wol})^a

ν	theory	δ_{exptl}^b	δ_{Wol}^c
1	4161.1661(9)	0.03	-0.001(1)
2	8087.0026(17)	0.07	0.000(2)
3	11782.3908(25)	0.03	-0.001(3)
4	15250.3656(32)	0.06	-0.002(3)
5	18491.9296(37)	0.01	-0.002(4)
6	21505.7795(43)	0.00	-0.004(4)
7	24287.9152(47)	0.01	-0.004(5)
8	26831.0906(50)	-0.07	-0.003(5)
9	29124.0404(52)	-0.05	-0.004(5)
10	31150.3909(51)	-0.08	-0.004(5)
11	32887.0983(49)	-0.03	-0.004(5)
12	34302.1740(42)	-0.03	-0.004(4)
13	35351.3145(31)	-0.05	-0.003(3)
14	35973.2731(16)	-0.11	0.000(2)

^aAll energies in cm^{-1} . ^bDabrowski¹⁹ estimates the uncertainty of the measurements as 0.1 cm^{-1} . ^cWolniewicz²⁶ does not give uncertainties for individual levels. He estimates overall accuracy as 0.001 cm^{-1} . The error estimation in this column comes only from the uncertainty of our calculations.

without including the Araki-Sucher term^{43,44} and the interatomic distance dependence of the two-electron Bethe logarithm.

In Table 4, we compare our predictions for the lowest rotational and vibrational excitation energies. Agreement with the experimental values for H_2 is excellent, but for the vibrational excitation in D_2 , a small discrepancy (of 2 experimental σ) is observed. Note that the experimental vibrational energies shown in this table are not measured directly but are extracted from a set of experimental literature lines fitted to the Dunham expansion.

Table 7. Comparison of the Theoretical and Experimental Quadrupole Transition Energies in H₂. δ is their difference and carries the combined experimental/theoretical uncertainty. All energies in cm⁻¹.

<i>J</i>	<i>S</i> ₀ (<i>J</i>)			<i>S</i> ₁ (<i>J</i>)		
	theory	experiment ⁵³	δ	theory	experiment ²⁹	δ
0	354.373 13(3)	354.373 50(40)	-0.000 37(40)		354.373 3(2)	-0.000 2(2)
1	587.032 02(4)	587.032 11(17)	-0.000 09(17)		587.032 1(3)	0.000 0(3)
2	814.424 30(6)	814.424 73(8)	-0.000 43(10)		814.424 9(2)	-0.000 6(2)
3	1 034.670 68(8)	1 034.670 24(3)	0.000 44(9)		1 034.670 6(3)	0.000 1(3)
4	1 246.099 54(9)	1 246.098 11(17)	0.001 43(19)		1 246.100(5)	0.000(5)
5	1 447.280 93(11)	1 447.278 82(41)	0.002 11(42)		1 447.282(5)	-0.002(5)

<i>J</i>	<i>Q</i> ₁ (<i>J</i>)			<i>S</i> ₁ (<i>J</i>)		
	theory	experiment ⁵⁴	δ	theory	experiment ⁵⁴	δ
0				4497.8384(9)	4497.8391(2)	-0.0007(9)
1	4155.2538(9)	4155.25469(8)	-0.0009(9)	4712.9046(9)	4712.9054(2)	-0.0008(9)
2	4143.4653(9)	4143.4660(3)	-0.0007(9)	4917.0063(10)	4917.0069(3)	-0.0006(10)
3	4125.8726(9)	4125.8739(4)	-0.0013(10)	5108.4029(10)	5108.4040(6)	-0.0011(12)
4	4102.5820(9)	4102.582(4)	0.000(4)			

<i>J</i>	<i>Q</i> ₂ (<i>J</i>)			<i>S</i> ₂ (<i>J</i>)		
	theory	experiment ⁵⁴	δ	theory	experiment ⁵⁴	δ
0				8406.3608(18)	8406.365(2)	-0.004(3)
1	8075.3074(17)	8075.3114(6)	-0.0040(18)	8604.2152(18)	8604.2189(8)	-0.0037(20)
2	8051.9877(17)	8051.991(7)	-0.003(7)	8785.5244(18)	8785.529(6)	-0.005(6)
3	8017.1831(17)	8017.19(1)	-0.01(1)			

Table 8. Comparison of the Theoretical and Experimental Quadrupole Transition Energies in D₂. δ is their difference and carries the combined experimental/theoretical uncertainty. All energies in cm⁻¹.

<i>J</i>	<i>S</i> ₀ (<i>J</i>)			<i>S</i> ₁ (<i>J</i>)		
	theory	experiment ⁵⁵	δ	theory	experiment ⁵⁶	δ
0	179.06710(1)	179.068(2)	-0.001(2)	3166.3605(2)	3166.3596(40)	0.0009(40)
1	297.53374(1)	297.533(3)	0.001(3)	3278.5231(2)	3278.5222(40)	0.0009(40)
2	414.64845(2)	414.648(2)	0.000(2)	3387.2626(2)	3387.2606(50)	0.0020(50)
3	529.90025(2)	529.900(4)	0.000(4)	3492.0937(2)	3492.0913(40)	0.0024(40)
4	642.80664(2)	642.806(4)	0.001(4)			
5	752.91993(3)	752.923(20)	-0.003(20)			
6	859.83237(3)	859.845(20)	-0.013(20)			

<i>J</i>	<i>O</i> ₁ (<i>J</i>)			<i>Q</i> ₁ (<i>J</i>)		
	theory	experiment ⁵⁶	δ	theory	experiment ⁵⁶	δ
1				2991.5070(2)	2991.5043(40)	0.0027(40)
2	2814.5500(2)	2814.5459(40)	0.0041(40)	2987.2934(2)	2987.2955(40)	-0.0021(40)
3	2693.9733(2)	2693.9723(40)	0.0010(40)	2980.9894(2)	2980.9882(40)	0.0012(40)
4	2572.645 0(2)	2572.6428(50)	0.0022(50)	2972.6141(2)	2972.6128(50)	0.0013(50)

The error estimation of the obtained value is one standard deviation of this fit.^{45,46} We can also compare our result for D₂ with recent fully nonadiabatic relativistic calculations through order α^2 (including all recoil terms of the order of α^2/μ_n) performed by Bubin et al.⁴⁵ For the energy difference of rotationless $\nu = 1$ and $\nu = 0$ states of D₂, these authors obtained

2993.6326(10) cm⁻¹, in perfect agreement with our value 2993.6326(2) cm⁻¹ obtained neglecting recoil terms. The authors of ref 45 attributed a discrepancy of their value with experimental results to the missing QED correction, unknown when their work was published. In fact, when our QED correction, amounting to -0.0155 cm⁻¹, is added, a small

disagreement of $0.004(2) \text{ cm}^{-1}$, the same as shown in Table 4, remains.

A detailed comparison of the theoretical rotational excitation energies with a very recent experimental determination²⁹ is presented in Table 5. The error estimation assigned to the theoretical energy differences has been computed in the same fashion as the uncertainty for the individual levels, described earlier. In general, the assumed theoretical uncertainties are much smaller than the experimental ones. Very good agreement between the theoretical and experimental values is observed for all measured levels up to $J = 16$, except for $J = 4$, where a 4 σ unexplained discrepancy is present.

In the two last columns of Table 5, we also present the pure QED contribution to each measured rotational level obtained by subtracting the nonrelativistic and our relativistic energies from either the total theoretical or the experimental values. The comparison of the theoretical and experimental QED contributions is shown in Figure 3. The observed agreement is remarkably good (except for $J = 4$), confirming the accuracy of both the measurement and the calculations. To our knowledge, this agreement represents the first observation of QED effects in a molecular spectrum. Specifically, these effects include the electron self-energy (interaction of the electron with the vacuum fluctuations of the electromagnetic field), the vacuum polarization (electron–positron virtual pair creation), and the retardation of the electron–electron interaction.⁹ The conventional relativistic quantum mechanics based on the Dirac–Coulomb (DC) or Dirac–Coulomb–Breit (DCB) equations³¹ neglects this complicated physics. Had we assumed the presence of exactly two electrons in the molecule and solved the DC or DCB equations using the direct perturbation theory (DPT) expansion in even powers of α ,^{47,48} then all theoretical points in Figure 3 would have lain very close to the straight line $E = 0$. The deviation from this line is a quantitative demonstration of QED contribution to the rotational excitation energies of H_2 . It may be pointed out in this context that due to the Brown–Ravenhall disease,⁴⁹ the eigenenergies of the DC or DCB equations are not precisely defined⁵⁰ and have an intrinsic uncertainty on the order of α^3 ,⁵¹ the same as the leading QED effects. Thus, when adding the α^3 QED contribution to relativistic energies obtained using methods other than DPT, one should take special care to avoid double counting.

Whereas the effects of QED are well seen in the rotational spectrum, their influence on the bond length in H_2 is very small. The rotational constant B , defined as half of the smallest rotational spacing, decreases only by about 8×10^{-6} in relative terms, which leads to an increase of the H–H bond distance (assumed to be proportional to $B^{-1/2}$) by merely $3 \times 10^{-6} \text{ \AA}$. One should note, however, that for very weakly bound dimers, like e.g. He_2 , the QED effects may significantly affect the bond length.⁵²

It appears that the experimental pure vibrational excitation energies ($J = 0$) have not been measured with accuracy comparable to that of pure rotational ones ($\nu = 0$) of ref 29. In Table 6, we show a comparison of our values with the experimental data of Dabrowski¹⁹ and with the theoretical values of Wolniewicz.²⁶ The agreement in both cases is very good, although it should be kept in mind that the accuracy of Dabrowski's results is limited. She estimates it as 0.1 cm^{-1} without referring to any specific part of the spectrum. We observed, however, that for some high-lying levels (i.e., the highest bound levels shown in Table 4 of ref 19), her energies differ from ours by $0.2\text{--}0.4 \text{ cm}^{-1}$, i.e., much more

than our estimated uncertainties. The excellent agreement between ours and Wolniewicz's results, seen in Table 6, is likely due to the fact that the contribution of the two-electron QED effects, neglected by Wolniewicz, is small for the presented vibrational energy differences.

Some highly accurate values of the Q and S branch quadrupole transition frequencies of H_2 and D_2 are available in the literature^{53–56} and are compared with our results in Tables 7 and 8. Both for H_2 and for D_2 we observe now some disagreements between theory and experimental results. The most significant discrepancies occur for the $S_0(2) - S_0(5)$ transitions in H_2 measured by Jennings and Brault.⁵³ Their transition energies differ from 5 up to 15 experimental σ from our values. The reason for this discrepancy is not obvious to us. We speculate that an explanation may come from a possible underestimation of higher-order nonadiabatic effects, which are larger for H_2 than D_2 . An alternative explanation, put forward for consideration by Campargue,⁵⁷ is a small pressure shift affecting the experimental result. On the other hand, we observe good agreement of our predictions with somewhat less accurate $S_0(J)$ transition energies that can be obtained from very recent measurement by Salumbides et al.²⁹ [except for the $S_0(2)$ line related to a similar exception noted in Table 5 for $J = 4$]. For the Q_1 and S_1 transitions, the discrepancies exceed the experimental error but fit within the sum of the theoretical and experimental uncertainties. Among the Q_2 and S_2 transitions, two lines ($J = 1$) do not conform to the above description and show discrepancies almost twice the size of the sum of these two uncertainties. Campargue,⁵⁷ on the basis of his own measurements and a comparison with Rahn and Rosasco's results,⁵⁸ suggests that the discrepancy in the Q_1 branch may be due to the calibration of the spectra of ref 54. What is more, we observe very good agreement of theoretical line positions with the measurements by Ferguson et al.⁵⁹ for higher transitions (not shown in Tables): $S_4(0)$, $S_4(1)$, and $S_5(1)$. The differences (and the experimental uncertainties) are $0.0008(36)$, $0.0001(10)$, and $0.0014(200) \text{ cm}^{-1}$, respectively.

For D_2 (see Table 8), we find general agreement with the measurements of Jennings et al.⁵⁵ and of McKellar and Oka.⁵⁶ The differences δ between the experimental and theoretical frequencies fit well within the experimental uncertainties of several thousandths of cm^{-1} . We observe, however, a discrepancy with recently reported measurements⁶⁰ of $S_1(0)$ and $S_1(1)$ lines. The experimental values $3166.3620(2) \text{ cm}^{-1}$ and $3278.5220(2) \text{ cm}^{-1}$ differ from our predictions by several combined experimental/theoretical uncertainties.

7. CONCLUSION

The accuracy of about 0.001 cm^{-1} for most of the dissociation energies of rovibrational levels of H_2 and isotopomers has been achieved due to the recent progress made in two directions. The first one enabled a complete treatment of the leading α^3 QED effects. In particular, an effective approach to calculating the many-electron Bethe logarithm and mean values of singular operators, like the Araki–Sucher term,^{43,44} has been developed and is given in refs 27, 61 and 62. The second direction, indispensable to reaching this accuracy, is the nonadiabatic perturbation theory, NAPT,^{6,7,63} which enables a rigorous approach to the finite nuclear mass effects beyond the adiabatic approximation. However, nonadiabatic contributions to the relativistic and QED energies (the so-called recoil corrections)

still remain to be evaluated. These corrections as well as the higher-order nonadiabatic $\mathcal{O}(\mu_n^{-5/2})$, the complete α^4 QED, and the leading term in α^5 correction have to be evaluated in order to reach a 10^{-6} cm^{-1} level of accuracy, where the H_2 spectrum becomes sensitive to uncertainties in the electron–proton mass ratio and in the proton charge radius.⁶⁴ One may speculate that at this level of precision exotic forces between hadrons carried by very weakly interacting sub-electronvolt particles (WISPs) might also become spectroscopically observable.⁶⁵

■ ASSOCIATED CONTENT

S Supporting Information. Components of the dissociation energy as well as its uncertainty for all the rovibrational levels of H_2 and D_2 . This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: komasa@man.poznan.pl

■ ACKNOWLEDGMENT

We wish to thank Wim Ubachs and Edcel Salumbides for useful discussions and comments on the manuscript. J.K. acknowledges support by the Polish Ministry of Science and Higher Education Grant No. N N204 015338 and by a computing grant from Poznań Supercomputing and Networking Center. B.J. acknowledges support by the grant N N204 182840. K.P. acknowledges support by NIST through Precision Measurement Grant PMG 60NANB7D6153.

■ REFERENCES

- (1) Liu, J.; Salumbides, E. J.; Hollenstein, U.; Koelemeij, J. C. J.; Eikema, K. S. E.; Ubachs, W.; Merkt, F. *J. Chem. Phys.* **2009**, *130*, 174306.
- (2) Liu, J.; Sprecher, D.; Jungen, C.; Ubachs, W.; Merkt, F. *J. Chem. Phys.* **2010**, *132*, 154301.
- (3) Sprecher, D.; Liu, J.; Jungen, C.; Ubachs, W.; Merkt, F. *J. Chem. Phys.* **2010**, *133*, 111102.
- (4) Bailly, D.; Salumbides, E. J.; Vervloet, M.; Ubachs, W. *Mol. Phys.* **2010**, *108*, 827.
- (5) Pachucki, K. *Phys. Rev. A* **2010**, *82*, 032509.
- (6) Pachucki, K.; Komasa, J. *J. Chem. Phys.* **2008**, *129*, 034102.
- (7) Pachucki, K.; Komasa, J. *J. Chem. Phys.* **2009**, *130*, 164113.
- (8) Bethe, H. A.; Salpeter, E. E. *Quantum Mechanics of One- and Two-Electron Systems*; Springer-Verlag: Berlin, 1957; p 181.
- (9) Sapirstein, J. In *Springer Handbook of Atomic, Molecular, and Optical Physics*; Drake, G. W. F., Ed.; Springer: New York, 2006; p 413.
- (10) Caswell, W. E.; Lepage, G. P. *Phys. Lett. B* **1986**, *167*, 437.
- (11) Pachucki, K. *Phys. Rev. A* **1997**, *56*, 297.
- (12) Pachucki, K.; Komasa, J. *Phys. Rev. Lett.* **2004**, *92*, 213001.
- (13) Yan, Z.-C.; Nörtershäuser, W.; Drake, G. W. F. *Phys. Rev. Lett.* **2008**, *100*, 243002.
- (14) Puchalski, M.; Pachucki, K. *Phys. Rev. A* **2008**, *78*, 052511.
- (15) Korobov, V. I. *Phys. Rev. A* **2008**, *77*, 022509.
- (16) Wolniewicz, L. *J. Chem. Phys.* **1983**, *78*, 6173.
- (17) Herzberg, G.; Howe, L. L. *Can. J. Phys.* **1959**, *37*, 636.
- (18) Kolos, W.; Szalewicz, K.; Monkhorst, H. J. *J. Chem. Phys.* **1986**, *84*, 3278.
- (19) Dabrowski, I. *Can. J. Phys.* **1984**, *62*, 1639.
- (20) Kolos, W.; Wolniewicz, L. *J. Chem. Phys.* **1964**, *41*, 3663.
- (21) James, H. M.; Coolidge, A. S. *J. Chem. Phys.* **1933**, *1*, 825.
- (22) Wolniewicz, L. *J. Chem. Phys.* **1993**, *99*, 1851.
- (23) Kolos, W.; Wolniewicz, L. *J. Chem. Phys.* **1966**, *45*, 509.
- (24) Schwartz, C.; LeRoy, R. J. *J. Mol. Spectrosc.* **1987**, *121*, 420.
- (25) Chuang, M.-C.; Zare, R. N. *J. Mol. Spectrosc.* **1987**, *121*, 380.
- (26) Wolniewicz, L. *J. Chem. Phys.* **1995**, *103*, 1792.
- (27) Piszczatowski, K.; Lach, G.; Przybytek, M.; Komasa, J.; Pachucki, K.; Jeziorski, B. *J. Chem. Theory Comput.* **2009**, *5*, 3039.
- (28) Pachucki, K.; Komasa, J. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9188.
- (29) Salumbides, E. J.; Dickenson, G. D.; Ivanov, T. I.; Ubachs, W. *Phys. Rev. Lett.* **2011**, *107*, 043005.
- (30) Pyykkö, P.; Dyal, K. G.; Császár, A. G.; Tarczay, G.; Polyansky, O. L.; Tennyson, J. *Phys. Rev. A* **2001**, *63*, 024502.
- (31) Pyykkö, P. *Chem. Rev.* **2011**. In press, DOI: 10.1021/cr200042e.
- (32) Herman, R. M.; Asgharian, A. *J. Mol. Spectrosc.* **1966**, *19*, 305.
- (33) Bunker, P.; McLarnon, C.; Moss, R. *Mol. Phys.* **1977**, *33*, 425.
- (34) Kutzelnigg, W. *Mol. Phys.* **2007**, *105*, 2627.
- (35) Jaquet, R.; Kutzelnigg, W. *Chem. Phys.* **2008**, *346*, 69.
- (36) Pachucki, K. *Phys. Rev. A* **2006**, *74*, 022512.
- (37) Pachucki, K. *Phys. Rev. A* **2006**, *74*, 062510.
- (38) Pachucki, K.; Komasa, J. *Phys. Rev. A* **2011**, *83*, 042510.
- (39) Pachucki, K. *Phys. Rev. A* **2009**, *80*, 032520.
- (40) Kolos, W.; Wolniewicz, L. *J. Chem. Phys.* **1966**, *45*, 944.
- (41) Piszczatowski, K.; Lach, G.; Jeziorski, B. *Phys. Rev. A* **2008**, *77*, 062514.
- (42) Sprecher, D.; Jungen, C.; Ubachs, W.; Merkt, F. *Faraday Discuss.* **2011**, *150*, 51.
- (43) Araki, H. *Prog. Theor. Phys.* **1957**, *17*, 619.
- (44) Sucher, J. *Phys. Rev.* **1958**, *109*, 1010.
- (45) Bubin, S.; Stanke, M.; Molski, M.; Adamowicz, L. *Chem. Phys. Lett.* **2010**, *494*, 21.
- (46) Stanke, M.; Kedziera, D.; Bubin, S.; Molski, M.; Adamowicz, L. *J. Chem. Phys.* **2008**, *128*, 114313.
- (47) Rutkowski, A. *J. Phys. B* **1986**, *19*, 149, 3431, 3443.
- (48) Kutzelnigg, W. *Z. Phys. D* **1989**, *11*, 15.
- (49) Brown, G. E.; Ravenhall, D. G. *Proc. R. Soc. London, Ser. A* **1951**, *208*, 552.
- (50) Kutzelnigg, W. *Chem. Phys.* **2011**. In press, DOI: 10.1016/j.chemphys.2011.06.001.
- (51) Pestka, G.; Bylicki, M.; Karwowski, J. *J. Phys. B* **2007**, *40*, 2249.
- (52) Przybytek, M.; Cencek, W.; Komasa, J.; Lach, G.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2010**, *104*, 183003.
- (53) Jennings, D. E.; Brault, J. W. *J. Mol. Spectrosc.* **1983**, *102*, 265.
- (54) Bragg, S. L.; Smith, W. H.; Brault, J. W. *Astrophys. J.* **1982**, *263*, 999.
- (55) Jennings, D. E.; Weber, A.; Brault, J. W. *Appl. Opt.* **1986**, *25*, 284.
- (56) McKellar, A. R. W.; Oka, T. *Can. J. Phys.* **1978**, *56*, 1315.
- (57) Campargue, A. 2011, private communication.
- (58) Rahn, L. A.; Rosasco, G. *J. Phys. Rev. A* **1990**, *41*, 3698.
- (59) Ferguson, D. W.; Rao, K. N.; Mickelson, M. E.; Larson, L. E. *J. Mol. Spectrosc.* **1993**, *160*, 315.
- (60) Maddaloni, P.; Malara, P.; De Tommasi, E.; De Rosa, M.; Ricciardi, I.; Gagliardi, G.; Tamassia, F.; Di Lonardo, G.; De Natale, P. *J. Chem. Phys.* **2010**, *133*, 154317.
- (61) Pachucki, K.; Komasa, J. *J. Chem. Phys.* **2006**, *124*, 064308.
- (62) Pachucki, K.; Cencek, W.; Komasa, J. *J. Chem. Phys.* **2005**, *122*, 184101.
- (63) Pachucki, K. *Phys. Rev. A* **2010**, *81*, 032505.
- (64) Pohl, A.; et al. *Nature* **2010**, *466*, 213.
- (65) Jaeckel, J.; Ringwald, A. *Ann. Rev. Nucl. Part. Sci.* **2010**, *60*, 405.
- (66) Eyler, E. E.; Melikechi, N. *Phys. Rev. A* **1993**, *48*, R18.
- (67) Zhang, Y. P.; Cheng, C. H.; Kim, J. T.; Stanojevic, J.; Eyler, E. E. *Phys. Rev. Lett.* **2004**, *92*, 203003.
- (68) Jennings, D. E.; Bragg, S. L.; Brault, J. W. *Astrophys. J.* **1984**, *282*, L85.

Correlation Energy Expressions from the Adiabatic-Connection Fluctuation–Dissipation Theorem Approach

János G. Ángyán,^{*,†} Ru-Fen Liu,^{†,||} Julien Toulouse,^{*,‡} and Georg Jansen^{*,§}

[†]CRM2, Institut Jean Barriol, Nancy University and CNRS, 54506 Vandoeuvre-lès-Nancy, France

[‡]Laboratoire de Chimie Théorique, Université Pierre et Marie Curie and CNRS, 75005 Paris, France

[§]Fakultät für Chemie, Universität Duisburg-Essen, 45117 Essen, Germany

ABSTRACT: We explore several random phase approximation (RPA) correlation energy variants within the adiabatic-connection fluctuation–dissipation theorem approach. These variants differ in the way the exchange interactions are treated. One of these variants, named dRPA-II, is original to this work and closely resembles the second-order screened exchange (SOSEX) method. We discuss and clarify the connections among different RPA formulations. We derive the spin-adapted forms of all the variants for closed-shell systems and test them on a few atomic and molecular systems with and without range separation of the electron–electron interaction.

1. INTRODUCTION

There is a recent revival of interest in the random phase approximation (RPA) to obtain ground-state correlation energies of electronic systems.^{1–41} The RPA is considered a promising first approximation to obtain nonperturbative, exact-exchange-compatible, post-Kohn–Sham correlation energy corrections in density-functional theory. In particular, the RPA is thought of as a remedy for the bad description of London dispersion forces by conventional local and semilocal density-functional approximations. However, it is widely admitted that while RPA is well adapted to long-range electron–electron interactions, for small interelectronic distances its performance is even poorer than that of semilocal density functionals.^{42,43} An efficient way to make optimal use of RPA is to apply it in a range-separated approach,^{44,45} where the short-range interactions are described by an exchange–correlation density functional, and long-range exchange and correlation are treated by Hartree–Fock (HF) and RPA, respectively. Computational schemes following these principles have been recently proposed and applied mainly to van der Waals complexes.^{15–17,28,31,33,46}

Several formulations of RPA have been developed. Perhaps, the most well-known approach to RPA is the one based on the adiabatic-connection fluctuation–dissipation theorem (ACFDT).^{47,48} In this approach, the correlation energy expression involves integrations over both the frequency and the interaction strength, which can be performed either numerically or analytically. Obviously, an expression which has already been integrated analytically along at least one or both of these variables is more advantageous than the repeated use of numerical quadratures. If an analytical integration over the frequency is performed first, followed by a numerical integration over the interaction strength, one obtains an expression that is of the form of an interaction-strength-averaged two-particle density matrix contracted with the two-electron integrals. This is the *adiabatic-connection formulation*. An analytical integration over the interaction strength followed by a numerical integration along the frequency leads to an expression involving the dynamic dielectric matrix. This is the *dielectric-matrix formulation*. With a second analytical integration

(either along the interaction strength starting from the adiabatic-connection expression or along the frequency starting from the dielectric-matrix expression) both of these intermediate forms can be converted to a common expression, which consists of a sum of the shifts of electronic excitation energies when passing from an independent-particle to the RPA description of the excited states. This is the *plasmon formulation*. The plasmon expression can be further converted to an equivalent expression involving coupled-cluster doubles (CCD) amplitudes calculated in the ring-diagram approximation.¹⁴ This is the *ring CCD formulation*. The relationship between the adiabatic-connection and ring CCD formulations of RPA has been recently discussed in ref 34.

In this work, we study different variants of RPA within the adiabatic-connection formulation, which differ in the way the exchange interactions are handled. If the exchange interactions are neglected in the density matrix, we obtain the direct RPA (dRPA) approach (also called time-dependent Hartree), while inclusion of the nonlocal HF exchange response kernel leads to the RPax approach (also called time-dependent Hartree–Fock, or full RPA). A third possibility, not discussed here, consists of including an exact exchange response kernel from a local exact exchange potential.²⁷ If the dRPA density matrix is contracted with nonantisymmetrized two-electron integrals, one obtains what we call the dRPA-I variant, while if it is contracted with antisymmetrized two-electron integrals, one obtains the dRPA-II variant. Similarly, if the RPax density matrix is contracted with nonantisymmetrized two-electron integrals, the RPax-I variant is obtained, while if it is contracted with antisymmetrized two-electron integrals, one obtains the RPax-II variant. The dRPA-I variant is just the commonly called “RPA” of the density-functional/material-science community. The dRPA-II variant, which is similar to the second-order screened exchange (SOSEX) expression introduced by Grüneis et al.²³ in the ring CCD formulation, is original to this work. In contrast to SOSEX, it

Received: July 20, 2011

Published: August 31, 2011

involves higher-order screened exchange effects. The RPAX-II variant was first introduced by McLachlan and Ball,⁴⁹ but here we derive a new adiabatic-connection expression for it. Finally, the RPAX-I variant has been recently introduced by Toulouse et al.^{15,33} When possible, for the case of dRPA-I and RPAX-II, we also compare with the equivalent plasmon formulation and clarify the origin of the prefactor of 1/4 in the plasmon formula of RPAX-II in place of the prefactor of 1/2 appearing for dRPA-I. We remind the reader that in spite of the very different formulations, the dRPA-I variant is the same as the direct ring-CCD method, while the RPAX-II approach is identical to ring-CCD.^{34,46}

For the sake of simplicity, we give all the expressions without range separation, but it is straightforward to generalize them for the case of range separation, as done in ref 33. The paper is organized as follows. In section 2, we first provide an overview of the adiabatic-connection RPA correlation energy variants. In section 3, we review how the two-particle density matrix is obtained from the RPA polarization propagator. In section 4, we derive the expressions of RPA correlation energy variants in a spin-orbital basis. In section 5, we derive the corresponding spin-adapted expressions for closed-shell systems. In section 6, we perform numerical comparisons of different variants on a few atomic and molecular systems with and without range separation. Finally, section 7 contains our conclusions. The analysis of the second-order limit in the electron-electron interaction of each variant is given in the Appendix.

2. OVERVIEW OF RPA CORRELATION ENERGY VARIANTS IN THE ADIABATIC-CONNECTION FORMULATION

In the adiabatic-connection formalism, the correlation energy in a spin-orbital basis can be expressed as

$$E_c = \frac{1}{2} \int_0^1 d\alpha \operatorname{Tr}\{\mathbb{V}P_{c,\alpha}\} = \frac{1}{2} \int_0^1 d\alpha \sum_{pq,rs} \langle rq|sp\rangle (P_{c,\alpha})_{pq,rs} \quad (1)$$

where $\mathbb{V}_{sr,qp} = \langle rq|sp\rangle$ are the two-electron integrals, $P_{c,\alpha}$ is the correlation part of the two-particle density matrix at interaction strength α , and Tr denotes the trace (sum over the indices r and s). Using the antisymmetry of $P_{c,\alpha}$ with respect to the permutation of the indices p and s , the correlation energy can also be expressed as

$$\begin{aligned} E_c &= \frac{1}{4} \int_0^1 d\alpha \operatorname{Tr}\{\mathbb{W}P_{c,\alpha}\} \\ &= \frac{1}{4} \int_0^1 d\alpha \sum_{pq,rs} \langle rq||sp\rangle (P_{c,\alpha})_{pq,rs} \end{aligned} \quad (2)$$

where $\mathbb{W}_{sr,qp} = \langle rq||sp\rangle = \langle rq|sp\rangle - \langle rq|ps\rangle$ are the antisymmetrized two-electron integrals. In RPA-type approximations, $P_{c,\alpha}$ is obtained via the fluctuation-dissipation theorem

$$P_{c,\alpha}^{\text{RPA}} = - \int_{-\infty}^{\infty} \frac{d\omega}{2\pi i} e^{i\omega 0^+} [\Pi_{\alpha}^{\text{RPA}}(\omega) - \Pi_0(\omega)] \quad (3)$$

where $\Pi_{\alpha}^{\text{RPA}}(\omega)$ is the four-index matrix representation of the dynamic polarization propagator at interaction strength α and frequency ω , and $\Pi_0(\omega)$ is the corresponding noninteracting

(Hartree-Fock or Kohn-Sham) polarization propagator. In the dRPA variant (or time-dependent Hartree), the polarization propagator is obtained from the response equation with the Hartree kernel \mathbb{V}

$$\Pi_{\alpha}^{\text{dRPA}}(\omega)^{-1} = \Pi_0(\omega)^{-1} - \alpha \mathbb{V} \quad (4)$$

whereas in the RPAX variant (or time-dependent Hartree-Fock), the polarization propagator is obtained using the Hartree-Fock kernel \mathbb{W}

$$\Pi_{\alpha}^{\text{RPAX}}(\omega)^{-1} = \Pi_0(\omega)^{-1} - \alpha \mathbb{W} \quad (5)$$

The obtained dRPA and RPAX correlation density matrices $P_{c,\alpha}^{\text{dRPA}}$ and $P_{c,\alpha}^{\text{RPAX}}$ are completely expressed in the basis of occupied-virtual orbital products, i.e., $pq = ia$ or ai and $rs = jb$ or bj where i and j refer to occupied orbitals and a and b to virtual orbitals. Neither $P_{c,\alpha}^{\text{dRPA}}$ nor $P_{c,\alpha}^{\text{RPAX}}$ are properly antisymmetric. As a consequence, the two correlation energy expressions, eqs 1 and 2, are no longer equivalent in dRPA or RPAX. This leads to at least four RPA variants for calculating correlation energies, denoted here by dRPA-I, dRPA-II, RPAX-I, and RPAX-II, depending on whether the correlation density matrix is contracted with the nonantisymmetrized two-electron integrals \mathbb{V} (variants I) or the antisymmetrized two-electron integrals \mathbb{W} (variants II).

The dRPA-I variant is obtained by inserting the dRPA correlation density matrix in eq 1:

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \int_0^1 d\alpha \operatorname{Tr}\{\mathbb{V}P_{c,\alpha}^{\text{dRPA}}\} \quad (6)$$

This variant is commonly called “RPA” in the density-functional/material-science community. It corresponds to the first RPA correlation energy approximation historically developed and is still widely used. Since the dRPA response equation involves the mere Hartree kernel, only the screening effect of the bare Coulomb interaction is taken into account in the polarization propagator and all exchange-correlation screening effects are neglected. The resulting correlation energies tend to be too strongly negative. At second order in the electron-electron interaction, the dRPA-I correlation energy does not reduce to the standard second-order Møller-Plesset (MP2) correlation energy but instead to a “direct MP2” expression, i.e., without the MP2 exchange term.^{2,50}

The dRPA-II variant is obtained by contracting the dRPA correlation density matrix with the antisymmetrized two-electron integrals \mathbb{W} :

$$E_c^{\text{dRPA-II}} = \frac{1}{2} \int_0^1 d\alpha \operatorname{Tr}\{\mathbb{W}P_{c,\alpha}^{\text{dRPA}}\} \quad (7)$$

which re-establishes the correct second-order MP2 limit. In view of equation 2, it could have been suggested to use a factor of 1/4 instead of 1/2 in eq 7, but in fact the correct MP2 limit is only recovered with the factor 1/2. This variant can also be obtained from eq 6 by antisymmetrizing the correlation density matrix with respect to the permutation of p and s : $(P_{c,\alpha}^{\text{dRPA}})_{pq,rs} \rightarrow (P_{c,\alpha}^{\text{dRPA}})_{pq,rs} - (P_{c,\alpha}^{\text{dRPA}})_{sq,rp}$. As far as we know, the dRPA-II variant has never been described before. It is similar to the second-order screened exchange (SOSEX) expression introduced by Grüneis et al.,²³ but the latter does not involve integration over the adiabatic connection and treats exchange effects only at the lowest order of perturbation.

The RPAX-I variant is obtained by inserting the RPAX correlation density matrix in eq 1:

$$E_c^{\text{RPAX-I}} = \frac{1}{2} \int_0^1 d\alpha \text{Tr}\{\mathbb{V}\mathbb{P}_{c,\alpha}^{\text{RPAX}}\} \quad (8)$$

and has been introduced recently by Toulouse et al.^{15,33} In this variant, the exchange screening effects are taken into account in the polarization propagator. The matrix $\mathbb{P}_{c,\alpha}^{\text{RPAX}}$ is properly antisymmetric at first order, and therefore the RPAX-I correlation energy has the correct MP2 limit. At higher orders, however, $\mathbb{P}_{c,\alpha}^{\text{RPAX}}$ violates antisymmetry properties to some extent.

The RPAX-II variant is obtained by inserting the RPAX correlation density matrix in eq 2:

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \int_0^1 d\alpha \text{Tr}\{\mathbb{W}\mathbb{P}_{c,\alpha}^{\text{RPAX}}\} \quad (9)$$

which can also be obtained from eq 8 by antisymmetrizing the correlation density matrix: $(\mathbb{P}_{c,\alpha}^{\text{RPAX}})_{pq,rs} \rightarrow (1/2)[(\mathbb{P}_{c,\alpha}^{\text{RPAX}})_{pq,rs} - (\mathbb{P}_{c,\alpha}^{\text{RPAX}})_{sq,rp}]$, the factor 1/2 being justified by the fact that $\mathbb{P}_{c,\alpha}^{\text{RPAX}}$ is already approximately antisymmetric, in contrast to $\mathbb{P}_{c,\alpha}^{\text{dRPA}}$. This variant was first introduced by McLachlan and Ball.⁴⁹ At second order, it properly reduces to MP2.

In the following, these four RPA correlation energy variants will be analyzed further, and working expressions will be given.

3. TWO-PARTICLE DENSITY MATRIX FROM THE POLARIZATION PROPAGATOR

We first briefly review how to extract a two-particle density matrix from the RPA polarization propagator. The noninteracting (Hartree–Fock or Kohn–Sham) polarization propagator $\Pi_0(\omega)$ writes

$$\Pi_0(\omega) = -(\Lambda_0 - \omega\Delta)^{-1} \quad (10)$$

where Λ_0 and Δ are 2×2 supermatrices

$$\Lambda_0 = \begin{pmatrix} \boldsymbol{\varepsilon} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varepsilon} \end{pmatrix} \text{ and } \Delta = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \quad (11)$$

each block being of dimension $N_o N_v \times N_o N_v$, where N_o and N_v are the numbers of occupied and virtual orbitals, respectively. The diagonal matrix $\boldsymbol{\varepsilon}$ contains the independent one-particle excitation energies, $\varepsilon_{ia,jb} = (\varepsilon_a - \varepsilon_i)\delta_{ij}\delta_{ab}$, and \mathbf{I} is the identity matrix. Similarly, the RPA polarization propagator at interaction strength α writes

$$\Pi_\alpha^{\text{RPA}}(\omega) = -(\Lambda_\alpha - \omega\Delta)^{-1} \quad (12)$$

where the supermatrix Λ_α is calculated with the Hartree kernel \mathbb{V} in the case of dRPA:

$$\Lambda_\alpha^{\text{dRPA}} = \Lambda_0 + \alpha\mathbb{V} \quad (13)$$

and with the Hartree–Fock kernel \mathbb{W} in the case of RPAX:

$$\Lambda_\alpha^{\text{RPAX}} = \Lambda_0 + \alpha\mathbb{W} \quad (14)$$

From now on, we will consider real-valued orbitals. In this case, the Hartree kernel is made of four identical blocks,

$$\mathbb{V} = \begin{pmatrix} \mathbf{K} & \mathbf{K} \\ \mathbf{K} & \mathbf{K} \end{pmatrix} \quad (15)$$

where $K_{ia,jb} = \langle ab|ij \rangle$ are nonantisymmetrized two-electron integrals. Similarly, the Hartree–Fock kernel writes

$$\mathbb{W} = \begin{pmatrix} \mathbf{A}' & \mathbf{B} \\ \mathbf{B} & \mathbf{A}' \end{pmatrix} \quad (16)$$

with the antisymmetrized two-electron integrals

$$A'_{ia,jb} = \langle ib||aj \rangle = \langle ib|aj \rangle - \langle ib|ja \rangle = K_{ia,jb} - J_{ia,jb} \quad (17)$$

and

$$B_{ia,jb} = \langle ab||ij \rangle = \langle ab|ij \rangle - \langle ab|ji \rangle = K_{ia,jb} - K'_{ia,jb} \quad (18)$$

Let us consider now the generalized nonhermitian RPA eigenvalue equation

$$\Lambda_\alpha C_{\alpha,n} = \omega_{\alpha,n} \Delta C_{\alpha,n} \quad (19)$$

whose solutions come in pairs: positive excitation energies $\omega_{\alpha,n}$ with eigenvectors $C_{\alpha,n} = (\mathbf{x}_{\alpha,n} \mathbf{y}_{\alpha,n})$ and negative excitation energies $\omega_{\alpha,-n} = -\omega_{\alpha,n}$ with eigenvectors $C_{\alpha,-n} = (\mathbf{y}_{\alpha,n} \mathbf{x}_{\alpha,n})$. The spectral representation of $\Pi_\alpha^{\text{RPA}}(\omega)$ then writes

$$\Pi_\alpha^{\text{RPA}}(\omega) = \sum_n \left\{ \frac{C_{\alpha,n} C_{\alpha,n}^T}{\omega - \omega_{\alpha,n} + i0^+} - \frac{C_{\alpha,-n} C_{\alpha,-n}^T}{\omega - \omega_{\alpha,-n} - i0^+} \right\} \quad (20)$$

where the sum is over eigenvectors n with positive excitation energies $\omega_{\alpha,n} > 0$. The fluctuation–dissipation theorem [eq 3] leads to the supermatrix representation of the correlation density matrix $\mathbb{P}_{c,\alpha}^{\text{RPA}}$ (using contour integration in the upper half of the complex plane):

$$\begin{aligned} \mathbb{P}_{c,\alpha}^{\text{RPA}} &= - \int_{-\infty}^{\infty} \frac{d\omega}{2\pi i} e^{i\omega 0^+} [\Pi_\alpha^{\text{RPA}}(\omega) - \Pi_0(\omega)] \\ &= \sum_n \{ C_{\alpha,-n} C_{\alpha,-n}^T - C_{0,-n} C_{0,-n}^T \} \end{aligned} \quad (21)$$

with the noninteracting eigenvectors $C_{0,-n} = (\mathbf{y}_{0,n} \mathbf{x}_{0,n})$ with $\mathbf{y}_{0,n} = \mathbf{0}$ and $\mathbf{x}_{0,n} = \mathbf{I}_n$ (where \mathbf{I}_n is the vector whose n th component is 1 and all other components are zero). The explicit supermatrix expression of the RPA correlation density matrix is thus

$$\mathbb{P}_{c,\alpha}^{\text{RPA}} = \begin{pmatrix} \mathbf{Y}_\alpha \mathbf{Y}_\alpha^T & \mathbf{Y}_\alpha \mathbf{X}_\alpha^T \\ \mathbf{X}_\alpha \mathbf{Y}_\alpha^T & \mathbf{X}_\alpha \mathbf{X}_\alpha^T \end{pmatrix} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \quad (22)$$

where \mathbf{X}_α and \mathbf{Y}_α are the matrices whose columns contain the eigenvectors $\mathbf{x}_{\alpha,n}$ and $\mathbf{y}_{\alpha,n}$. The dRPA and RPAX correlation density matrices have the same form in terms of the eigenvector matrices \mathbf{X}_α and \mathbf{Y}_α although the eigenvectors are of course different for dRPA and RPAX.

4. CORRELATION ENERGY EXPRESSIONS IN A SPIN–ORBITAL BASIS

We give here the expressions in a spin–orbital basis for calculating the different RPA correlation energy variants. We first consider the dRPA-I and RPAX-II variants which have similar expressions. In both cases, the integration over the adiabatic connection can be done analytically, leading to plasmon formulas. We then examine the dRPA-II and RPAX-I variants. They have in common that they mix the nonantisymmetrized integrals \mathbb{V} and the antisymmetrized integrals \mathbb{W} , which makes

it impossible to do the integration over the adiabatic connection analytically. Although the dRPA-I variant is well-documented in the literature after the work of Furche and co-workers,^{2,13,32} the review that we give here is useful to define our notations and for comparisons with other variants. The RPAX-I variant has been discussed in detail in the context of range separation by Toulouse et al.^{15,31,33} The RPAX-II variant is much less documented, and the dRPA-II is new, so most of the expressions that we give for them are original to this work.

4.1. dRPA-I Correlation Energy. There are several equivalent expressions for the dRPA-I correlation energy.

4.1.1. Adiabatic-Connection Formula. The dRPA-I correlation energy of eq 6 can be expressed with the eigenvectors of the dRPA polarization propagator according to the general prescription to form the correlation density matrix, eq 21:

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \int_0^1 d\alpha \sum_n \text{Tr} \{ \mathbf{V} \mathbf{C}_{\alpha,-n} \mathbf{C}_{\alpha,-n}^T - \mathbf{V} \mathbf{C}_{0,-n} \mathbf{C}_{0,-n}^T \} \quad (23)$$

or, using the explicit expressions in terms of the block matrix components [eqs 15 and 22]:

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \{ [(\mathbf{X}_\alpha + \mathbf{Y}_\alpha)(\mathbf{X}_\alpha + \mathbf{Y}_\alpha)^T - \mathbf{I}] \mathbf{K} \} \quad (24)$$

where tr refers to the trace now applied to the block matrices (which are half the size of the supermatrices). As shown by Furche,² one does not need to calculate explicitly the eigenvector matrices \mathbf{X}_α and \mathbf{Y}_α to get the correlation energy; it is sufficient to form the matrix

$$\mathbf{Q}_\alpha = (\mathbf{X}_\alpha + \mathbf{Y}_\alpha)(\mathbf{X}_\alpha + \mathbf{Y}_\alpha)^T \quad (25)$$

which can be obtained directly from the matrices involved in the RPA response equation. In the case of dRPA, it simply reads

$$\mathbf{Q}_\alpha^{\text{dRPA}} = \boldsymbol{\varepsilon}^{1/2} (\mathbf{M}_\alpha^{\text{dRPA}})^{-1/2} \boldsymbol{\varepsilon}^{1/2} \quad (26)$$

with

$$\mathbf{M}_\alpha^{\text{dRPA}} = \boldsymbol{\varepsilon}^{1/2} (\boldsymbol{\varepsilon} + 2\alpha \mathbf{K}) \boldsymbol{\varepsilon}^{1/2} \quad (27)$$

The adiabatic-connection formula for the dRPA-I correlation energy is then finally

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \{ [\mathbf{Q}_\alpha^{\text{dRPA}} - \mathbf{I}] \mathbf{K} \} \quad (28)$$

In previous papers, this equation was written with the matrix $\mathbf{P}_\alpha^{\text{dRPA}} = \mathbf{Q}_\alpha^{\text{dRPA}} - \mathbf{I}$.

4.1.2. Plasmon Formula. The plasmon formula for the dRPA-I correlation energy is found by starting from an equivalent form of eq 23:

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \int_0^1 d\alpha \sum_n \text{Tr} \{ \mathbf{C}_{\alpha,-n}^T \mathbf{V} \mathbf{C}_{\alpha,-n} - \mathbf{C}_{0,-n}^T \mathbf{V} \mathbf{C}_{0,-n} \} \quad (29)$$

obtained by a cyclic permutation of the matrices in the trace. Since the positive excitation energies can be written as^{13,49}

$$\omega_{\alpha,n}^{\text{dRPA}} = \mathbf{C}_{\alpha,-n}^T \boldsymbol{\Lambda}_\alpha^{\text{dRPA}} \mathbf{C}_{\alpha,-n} \quad (30)$$

the derivative of $\omega_{\alpha,n}$ with respect to α gives

$$\frac{d\omega_{\alpha,n}^{\text{dRPA}}}{d\alpha} = \mathbf{C}_{\alpha,-n}^T \frac{d\boldsymbol{\Lambda}_\alpha^{\text{dRPA}}}{d\alpha} \mathbf{C}_{\alpha,-n} = \mathbf{C}_{\alpha,-n}^T \mathbf{V} \mathbf{C}_{\alpha,-n} \quad (31)$$

which allows one to perform the integral over α in eq 29 analytically, leading to the plasmon formula

$$\begin{aligned} E_c^{\text{dRPA-I}} &= \frac{1}{2} \sum_n (\omega_{1,n}^{\text{dRPA}} - \omega_{0,n} - \mathbf{C}_{0,-n}^T \mathbf{V} \mathbf{C}_{0,-n}) \\ &= \frac{1}{2} \sum_n (\omega_{1,n}^{\text{dRPA}} - \omega_n^{\text{dTDA}}) \end{aligned} \quad (32)$$

where $\sum_n \omega_n^{\text{dTDA}} = \sum_n \mathbf{C}_{\alpha,-n}^T \boldsymbol{\Lambda}_1^{\text{dRPA}} \mathbf{C}_{\alpha,-n} = \sum_n \omega_{0,n} + \mathbf{C}_{0,-n}^T \mathbf{V} \mathbf{C}_{0,-n}$ is the sum of the (positive) excitation energies in the direct Tamm-Dancoff approximation (dTDA). The sum of the dTDA excitation energies can also be expressed as $\sum_n \omega_n^{\text{dTDA}} = \text{tr} \{ \boldsymbol{\varepsilon} + \mathbf{K} \}$.

4.1.3. Alternative Plasmon Formula. An alternative form of the plasmon formula can be found by rewriting eq 32 as

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \sum_n \text{Tr} \{ \boldsymbol{\Lambda}_1^{\text{dRPA}} \mathbf{C}_{1,-n} \mathbf{C}_{1,-n}^T - \boldsymbol{\Lambda}_1^{\text{dRPA}} \mathbf{C}_{0,-n} \mathbf{C}_{0,-n}^T \} \quad (33)$$

where the cyclic invariance of the trace has again been used. Using then eq 22 and recalling that the diagonal blocks of $\boldsymbol{\Lambda}_1^{\text{dRPA}}$ are $\boldsymbol{\varepsilon} + \mathbf{K}$ and the off-diagonal blocks are \mathbf{K} , the correlation energy becomes

$$\begin{aligned} E_c^{\text{dRPA-I}} &= \frac{1}{2} \text{tr} \{ [\mathbf{Y}_1 \mathbf{Y}_1^T + \mathbf{X}_1 \mathbf{X}_1^T - \mathbf{I}] (\boldsymbol{\varepsilon} + \mathbf{K}) \\ &\quad + [\mathbf{Y}_1 \mathbf{X}_1^T + \mathbf{X}_1 \mathbf{Y}_1^T] \mathbf{K} \} \end{aligned} \quad (34)$$

Introducing now the inverse of the \mathbf{Q}_α matrix:⁴³

$$\mathbf{Q}_\alpha^{-1} = (\mathbf{X}_\alpha - \mathbf{Y}_\alpha)(\mathbf{X}_\alpha - \mathbf{Y}_\alpha)^T \quad (35)$$

which in the case of dRPA can be written as

$$(\mathbf{Q}_\alpha^{\text{dRPA}})^{-1} = \boldsymbol{\varepsilon}^{-1/2} (\mathbf{M}_\alpha^{\text{dRPA}})^{1/2} \boldsymbol{\varepsilon}^{-1/2} \quad (36)$$

the correlation energy can be expressed as

$$\begin{aligned} E_c^{\text{dRPA-I}} &= \frac{1}{2} \text{tr} \left\{ \left[\frac{1}{2} (\mathbf{Q}_1^{\text{dRPA}} + (\mathbf{Q}_1^{\text{dRPA}})^{-1}) - \mathbf{I} \right] (\boldsymbol{\varepsilon} + \mathbf{K}) \right. \\ &\quad \left. + \frac{1}{2} (\mathbf{Q}_1^{\text{dRPA}} - (\mathbf{Q}_1^{\text{dRPA}})^{-1}) \mathbf{K} \right\} \end{aligned} \quad (37)$$

or, equivalently,

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \text{tr} \left\{ [\mathbf{Q}_1^{\text{dRPA}} - \mathbf{I}] \mathbf{K} + \frac{1}{2} [\mathbf{Q}_1^{\text{dRPA}} + (\mathbf{Q}_1^{\text{dRPA}})^{-1} - 2\mathbf{I}] \boldsymbol{\varepsilon} \right\} \quad (38)$$

or, rearranged in a different way

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \text{tr} \left\{ \frac{1}{2} \mathbf{Q}_1^{\text{dRPA}} (\boldsymbol{\varepsilon} + 2\mathbf{K}) + \frac{1}{2} (\mathbf{Q}_1^{\text{dRPA}})^{-1} \boldsymbol{\varepsilon} - (\boldsymbol{\varepsilon} + \mathbf{K}) \right\} \quad (39)$$

Using the expressions of $\mathbf{Q}_1^{\text{dRPA}}$ [eq 26], $(\mathbf{Q}_1^{\text{dRPA}})^{-1}$ [eq 36], and $\mathbf{M}_1^{\text{dRPA}}$ [eq 27] and the cyclic invariance of the trace, we finally arrive at the alternative form of the plasmon formula for

the dRPA-I correlation energy:

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \text{tr}\{(\mathbf{M}_1^{\text{dRPA}})^{1/2} - (\boldsymbol{\varepsilon} + \mathbf{K})\} \quad (40)$$

Recently, eq 40 was used by Eshuis et al.³² as the starting point for developing a computationally efficient algorithm for calculating the dRPA-I correlation energy. Note that expression 40 could have also been found by noting that the eigenvalues of $\mathbf{M}_1^{\text{dRPA}}$ are $(\omega_{1,n}^{\text{dRPA}})^2$ and, thus $\sum_n \omega_{1,n}^{\text{dRPA}} = \text{tr}\{(\mathbf{M}_1^{\text{dRPA}})^{1/2}\}$. However, working with \mathbf{Q}_α^{-1} will be useful for the other variants. Also, a comparison of eqs 28 and 38 provides us with a decomposition of the correlation energy into kinetic and potential contributions, $E_c^{\text{dRPA-I}} = T_c^{\text{dRPA-I}} + U_c^{\text{dRPA-I}}$. Indeed, the potential correlation energy is just the value of the integrand in eq 28 at $\alpha = 1$, i.e.,

$$U_c^{\text{dRPA-I}} = \frac{1}{2} \text{tr}\{[\mathbf{Q}_1^{\text{dRPA}} - \mathbf{I}]\mathbf{K}\} \quad (41)$$

and thus, by subtraction, according to eq 38, the kinetic correlation energy is

$$T_c^{\text{dRPA-I}} = \frac{1}{4} \text{tr}\{[\mathbf{Q}_1^{\text{dRPA}} + (\mathbf{Q}_1^{\text{dRPA}})^{-1} - 2\mathbf{I}]\boldsymbol{\varepsilon}\} \quad (42)$$

In the limit of a system with orbitals that are all degenerate, i.e., with static correlation only, then $\boldsymbol{\varepsilon} = \mathbf{0}$ and the kinetic correlation energy vanishes as it should. This is in agreement with the statement that dRPA-I correctly describes left–right static correlation in bond dissociations.^{7,51}

4.2. RPAX-II Correlation Energy. We now derive several equivalent RPAX-II correlation energy expressions by proceeding in an analogous way to the case of dRPA-I.

4.2.1. Adiabatic-Connection Formula. The RPAX-II correlation energy of eq 9 can be written in terms of the eigenvectors of the RPAX polarization propagator

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \int_0^1 d\alpha \sum_n \text{Tr}\{\mathbb{W}C_{\alpha,-n}C_{\alpha,-n}^T - \mathbb{W}C_{0,-n}C_{0,-n}^T\} \quad (43)$$

or, using the block structure of \mathbb{W} [eq 16]:

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \int_0^1 d\alpha \text{tr}\{(\mathbf{Y}_\alpha \mathbf{Y}_\alpha^T + \mathbf{X}_\alpha \mathbf{X}_\alpha^T - \mathbf{I})\mathbf{A}' + (\mathbf{Y}_\alpha \mathbf{X}_\alpha^T + \mathbf{X}_\alpha \mathbf{Y}_\alpha^T)\mathbf{B}\} \quad (44)$$

Using the matrix \mathbf{Q}_ω which in the case of RPAX is given by

$$\mathbf{Q}_\alpha^{\text{RPAX}} = (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' - \alpha\mathbf{B})^{1/2} (\mathbf{M}_\alpha^{\text{RPAX}})^{-1/2} (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' - \alpha\mathbf{B})^{1/2} \quad (45)$$

with

$$\mathbf{M}_\alpha^{\text{RPAX}} = (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' - \alpha\mathbf{B})^{1/2} (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' + \alpha\mathbf{B}) \times (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' - \alpha\mathbf{B})^{1/2} \quad (46)$$

and the inverse \mathbf{Q}_α^{-1}

$$(\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} = (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' - \alpha\mathbf{B})^{-1/2} (\mathbf{M}_\alpha^{\text{RPAX}})^{1/2} \times (\boldsymbol{\varepsilon} + \alpha\mathbf{A}' - \alpha\mathbf{B})^{-1/2} \quad (47)$$

we arrive at the adiabatic-connection formula for the RPAX-II correlation energy

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \int_0^1 d\alpha \text{tr}\left\{\frac{1}{2}\mathbf{Q}_\alpha^{\text{RPAX}}(\mathbf{A}' + \mathbf{B}) + \frac{1}{2}(\mathbf{Q}_\alpha^{\text{RPAX}})^{-1}(\mathbf{A}' - \mathbf{B}) - \mathbf{A}'\right\} \quad (48)$$

Since $\mathbf{Q}_\alpha = \mathbf{I} + \mathbf{P}_\alpha$ if \mathbf{P}_α is small, we can consider the approximation $\mathbf{Q}_\alpha^{-1} = (\mathbf{I} + \mathbf{P}_\alpha)^{-1} \approx \mathbf{I} - \mathbf{P}_\alpha = 2\mathbf{I} - \mathbf{Q}_\omega$ which leads to the following approximation for the RPAX-II correlation energy:

$$E_c^{\text{RPAX-IIa}} = \frac{1}{4} \int_0^1 d\alpha \text{tr}\left\{\frac{1}{2}\mathbf{Q}_\alpha^{\text{RPAX}}(\mathbf{A}' + \mathbf{B}) + \frac{1}{2}(2\mathbf{I} - \mathbf{Q}_\alpha^{\text{RPAX}})(\mathbf{A}' - \mathbf{B}) - \mathbf{A}'\right\} = \frac{1}{4} \int_0^1 d\alpha \text{tr}\{[\mathbf{Q}_\alpha^{\text{RPAX}} - \mathbf{I}]\mathbf{B}\} \quad (49)$$

So, we have the interesting result that this approximate correlation energy expression is analogous to the dRPA-I correlation energy expression of eq 28, the only differences being that matrix \mathbf{Q}_α is now obtained from the RPAX response equation and that it is contracted with the antisymmetrized two-electron integrals \mathbf{B} , along with the corresponding change of the prefactor from 1/2 to 1/4.

4.2.2. Plasmon Formula. As in the case of dRPA-I, the plasmon formula for the RPAX-II correlation energy is found by taking profit of the cyclic invariance of the trace to rewrite eq 43 as

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \int_0^1 d\alpha \sum_n \text{Tr}\{C_{\alpha,-n}^T \mathbb{W}C_{\alpha,-n} - C_{0,-n}^T \mathbb{W}C_{0,-n}\} \quad (50)$$

and then using $d\omega_{\alpha,n}^{\text{RPAX}}/d\alpha = C_{\alpha,-n}^T (d\Lambda_\alpha^{\text{RPAX}}/d\alpha) C_{\alpha,-n} = C_{\alpha,-n}^T \mathbb{W}C_{\alpha,-n}$ to integrate analytically over α

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \sum_n (\omega_{1,n}^{\text{RPAX}} - \omega_{0,n} - C_{0,-n}^T \mathbb{W}C_{0,-n}) = \frac{1}{4} \sum_n (\omega_{1,n}^{\text{RPAX}} - \omega_n^{\text{TDAX}}) \quad (51)$$

where $\sum_n \omega_n^{\text{TDAX}} = \sum_n C_{0,-n}^T \Lambda_1^{\text{RPAX}} C_{0,-n} = \sum_n \omega_{0,n} + C_{0,-n}^T \mathbb{W}C_{0,-n}$ is the sum of the (positive) excitation energies in the Tamm-Dancoff approximation with exchange (TDAX) or configuration interaction singles (CIS). The sum of the TDAX excitation energies can also be expressed as $\sum_n \omega_n^{\text{TDAX}} = \text{tr}\{\boldsymbol{\varepsilon} + \mathbf{A}'\}$. This plasmon formula was first presented by McLachlan and Ball.⁴⁹ The presence of a factor of 1/4 in eq 51 and not a factor of 1/2 like in eq 32 has been debated in the literature.⁵² The present exposition makes it clear that this factor of 1/4 is due to the use of the antisymmetrized two-electron integrals \mathbb{W} .

4.2.3. Alternative Plasmon Formula. As in the case of dRPA-I, the alternative plasmon formula is found by rewriting eq 51 as

$$E_c^{\text{RPAX-II}} = \frac{1}{4} \sum_n \text{Tr}\{\Lambda_1^{\text{RPAX}} C_{1,-n} C_{1,-n}^T - \Lambda_1^{\text{RPAX}} C_{0,-n} C_{0,-n}^T\} \quad (52)$$

and inserting the diagonal blocks of Λ_1^{RPax} which are $\boldsymbol{\varepsilon} + \mathbf{A}'$ and the off-diagonal blocks which are \mathbf{B} :

$$E_c^{\text{RPax-II}} = \frac{1}{4} \text{tr} \left\{ \frac{1}{2} \mathbf{Q}_1^{\text{RPax}} (\boldsymbol{\varepsilon} + \mathbf{A}' + \mathbf{B}) + \frac{1}{2} (\mathbf{Q}_1^{\text{RPax}})^{-1} (\boldsymbol{\varepsilon} + \mathbf{A}' - \mathbf{B}) - (\boldsymbol{\varepsilon} + \mathbf{A}') \right\} \quad (53)$$

Using the expressions of $\mathbf{Q}_1^{\text{RPax}}$ [eq 45], $(\mathbf{Q}_1^{\text{RPax}})^{-1}$ [eq 47], and $\mathbf{M}_1^{\text{RPax}}$ [eq 46] and the cyclic invariance of the trace, we arrive at the alternative plasmon formula for the RPax-II correlation energy:

$$E_c^{\text{RPax-II}} = \frac{1}{4} \text{tr} \{ (\mathbf{M}_1^{\text{RPax}})^{1/2} - (\boldsymbol{\varepsilon} + \mathbf{A}') \} \quad (54)$$

Finally, just as for dRPA-I, a comparison of eq 48 and eq 53 provides us with a decomposition of the correlation energy into the potential energy contribution to the correlation energy

$$U_c^{\text{RPax-II}} = \frac{1}{4} \text{tr} \left\{ \frac{1}{2} \mathbf{Q}_1^{\text{RPax}} (\mathbf{A}' + \mathbf{B}) + \frac{1}{2} (\mathbf{Q}_1^{\text{RPax}})^{-1} (\mathbf{A}' - \mathbf{B}) - \mathbf{A}' \right\} \quad (55)$$

and the kinetic correlation energy

$$T_c^{\text{RPax-II}} = \frac{1}{8} \text{tr} \{ [\mathbf{Q}_1^{\text{RPax}} + (\mathbf{Q}_1^{\text{RPax}})^{-1} - 2\mathbf{I}] \boldsymbol{\varepsilon} \} \quad (56)$$

The RPax-II kinetic correlation energy vanishes in the limit where $\boldsymbol{\varepsilon} = \mathbf{0}$ as for dRPA-I.

4.3. dRPA-II Correlation Energy. The dRPA-II correlation energy of eq 7 writes in terms of the eigenvectors of the dRPA polarization propagator

$$E_c^{\text{dRPA-II}} = \frac{1}{2} \int_0^1 d\alpha \sum_n \text{Tr} \{ \mathbb{W} \mathbf{C}_{\alpha,-n} \mathbf{C}_{\alpha,-n}^T - \mathbb{W} \mathbf{C}_{0,-n} \mathbf{C}_{0,-n}^T \} \quad (57)$$

leading to

$$E_c^{\text{dRPA-II}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \left\{ \frac{1}{2} \mathbf{Q}_\alpha^{\text{dRPA}} (\mathbf{A}' + \mathbf{B}) + \frac{1}{2} (\mathbf{Q}_\alpha^{\text{dRPA}})^{-1} (\mathbf{A}' - \mathbf{B}) - \mathbf{A}' \right\} \quad (58)$$

Equation 58 is similar to eq 48, with $\mathbf{Q}_\alpha^{\text{dRPA}}$ instead of $\mathbf{Q}_\alpha^{\text{RPax}}$ and a factor 1/2 instead of 1/4.

The approximation $\mathbf{Q}_\alpha^{-1} \approx 2\mathbf{I} - \mathbf{Q}_\alpha$ leads to the following approximate dRPA-II correlation energy:

$$E_c^{\text{dRPA-IIa}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \{ [\mathbf{Q}_\alpha^{\text{dRPA}} - \mathbf{I}] \mathbf{B} \} \quad (59)$$

which is in close analogy (but usually not equal) to the SOSEX correlation energy in the ring-CCD formulation. The analytic relationship of this ‘‘adiabatic-connection SOSEX’’ (AC-SOSEX) variant with the original SOSEX has been discussed in detail in ref 34.

4.4. RPax-I Correlation Energy. Finally, the RPax-I correlation energy of eq 8 writes in terms of the eigenvectors of the

RPax polarization propagator

$$E_c^{\text{RPax-I}} = \frac{1}{2} \int_0^1 d\alpha \sum_n \text{Tr} \{ \mathbb{V} \mathbf{C}_{\alpha,-n} \mathbf{C}_{\alpha,-n}^T - \mathbb{V} \mathbf{C}_{0,-n} \mathbf{C}_{0,-n}^T \} \quad (60)$$

leading to

$$E_c^{\text{RPax-I}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \{ [\mathbf{Q}_\alpha^{\text{RPax}} - \mathbf{I}] \mathbf{K} \} \quad (61)$$

which has the same form as eq 28 but with the RPax matrix $\mathbf{Q}_\alpha^{\text{RPax}}$. This last variant has been discussed in detail and applied in the context of range-separated density-functional theory.^{15,31,33}

5. CORRELATION ENERGY EXPRESSIONS IN A SPATIAL-ORBITAL BASIS FOR CLOSED-SHELL SYSTEMS

For spin-restricted closed-shell calculations, all of the matrices in the spin-orbital excitation basis occurring in the RPA equations have the following spin block structure:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\uparrow\uparrow,\uparrow\uparrow} & \mathbf{C}_{\uparrow\uparrow,\downarrow\downarrow} & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_{\downarrow\downarrow,\uparrow\uparrow} & \mathbf{C}_{\downarrow\downarrow,\downarrow\downarrow} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_{\uparrow\downarrow,\uparrow\downarrow} & \mathbf{C}_{\uparrow\downarrow,\uparrow\downarrow} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_{\downarrow\uparrow,\downarrow\uparrow} & \mathbf{C}_{\downarrow\uparrow,\downarrow\uparrow} \end{pmatrix} \quad (62)$$

This structure is a consequence of the fact that the two-electron integrals can be nonzero only for pairs of identical spins. The orthogonal transformation

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & -\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & -\mathbf{1} \end{pmatrix} \quad (63)$$

leads to a spin-adapted matrix $\tilde{\mathbf{C}} = \mathbf{U}^T \mathbf{C} \mathbf{U}$, which in the case of the matrices involved in RPA simplifies into a block-diagonal form with a spin-singlet excitation block ${}^1\mathbf{C}$ and three spin-triplet excitation blocks ${}^{3,0}\mathbf{C}$, ${}^{3,1}\mathbf{C}$, and ${}^{3,-1}\mathbf{C}$

$$\tilde{\mathbf{C}} = \begin{pmatrix} {}^1\mathbf{C} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {}^{3,0}\mathbf{C} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & {}^{3,1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & {}^{3,-1}\mathbf{C} \end{pmatrix} \quad (64)$$

with the matrix elements (i,j and a,b referring now to occupied and virtual spatial orbitals, respectively)

$${}^1\mathbf{C}_{ia,jb} = \frac{1}{2} (\mathbf{C}_{i\uparrow a\uparrow j\uparrow b\uparrow} + \mathbf{C}_{i\uparrow a\uparrow j\downarrow b\downarrow} + \mathbf{C}_{i\downarrow a\downarrow j\uparrow b\uparrow} + \mathbf{C}_{i\downarrow a\downarrow j\downarrow b\downarrow}) \quad (65a)$$

$${}^{3,0}\mathbf{C}_{ia,jb} = \frac{1}{2} (\mathbf{C}_{i\uparrow a\uparrow j\uparrow b\uparrow} - \mathbf{C}_{i\uparrow a\uparrow j\downarrow b\downarrow} - \mathbf{C}_{i\downarrow a\downarrow j\uparrow b\uparrow} + \mathbf{C}_{i\downarrow a\downarrow j\downarrow b\downarrow}) \quad (65b)$$

$${}^3, \pm 1 C_{ia,jb} = \frac{1}{2} (C_{i \uparrow a \downarrow j \uparrow b \downarrow} \pm C_{i \uparrow a \downarrow j \downarrow b \uparrow} \pm C_{i \downarrow a \uparrow j \uparrow b \downarrow} + C_{i \downarrow a \uparrow j \downarrow b \uparrow}) \quad (65c)$$

Let us start with dRPA. Spin-adaptation of the nonantisymmetrized integrals matrix \mathbf{K} gives only a contribution from the singlet excitations

$$\tilde{\mathbf{K}} = \begin{pmatrix} {}^1\mathbf{K} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \quad (66)$$

where ${}^1K_{ia,jb} = 2\langle ab|ij\rangle$. By eq 27, it leads to the following spin-adaptation for the matrix $\mathbf{M}_\alpha^{\text{dRPA}}$:

$$\tilde{\mathbf{M}}_\alpha^{\text{dRPA}} = \begin{pmatrix} {}^1\mathbf{M}_\alpha^{\text{dRPA}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varepsilon}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\varepsilon}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\varepsilon}^2 \end{pmatrix} \quad (67)$$

where ${}^1\mathbf{M}_\alpha^{\text{dRPA}} = \boldsymbol{\varepsilon}^{1/2}(\boldsymbol{\varepsilon} + 2\alpha{}^1\mathbf{K})\boldsymbol{\varepsilon}^{1/2}$, and $\boldsymbol{\varepsilon}$ refers now to the matrix of one-particle excitation energies indexed in spatial orbitals. By eq 26, it gives the following spin-adaptation for the matrix $\mathbf{Q}_\alpha^{\text{dRPA}}$:

$$\tilde{\mathbf{Q}}_\alpha^{\text{dRPA}} = \begin{pmatrix} {}^1\mathbf{Q}_\alpha^{\text{dRPA}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \quad (68)$$

where ${}^1\mathbf{Q}_\alpha^{\text{dRPA}} = \boldsymbol{\varepsilon}^{1/2}({}^1\mathbf{M}_\alpha^{\text{dRPA}})^{-1/2}\boldsymbol{\varepsilon}^{1/2}$.

Let us now consider RPax. Spin-adaptation of the antisymmetrized integrals matrices \mathbf{A}' and \mathbf{B} gives contributions from both singlet and triplet excitations:

$$\mathbf{A}' = \begin{pmatrix} {}^1\mathbf{A}' & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {}^3\mathbf{A}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & {}^3\mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & {}^3\mathbf{A}' \end{pmatrix}, \quad \tilde{\mathbf{B}} = \begin{pmatrix} {}^1\mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {}^3\mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & {}^3\mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -{}^3\mathbf{B} \end{pmatrix} \quad (69)$$

where ${}^1A'_{ia,jb} = 2\langle ib|aj\rangle - \langle ib|ja\rangle$, ${}^3A'_{ia,jb} = -\langle ib|ja\rangle$, ${}^1B_{ia,jb} = 2\langle ab|ij\rangle - \langle ab|ji\rangle$, and ${}^3B_{ia,jb} = -\langle ab|ji\rangle$. Notice the minus sign for the last triplet block in the $\tilde{\mathbf{B}}$ matrix, which makes spin-adaptation less trivial for RPax. By eq 46, it leads to the following

spin-adaptation for the matrix $\mathbf{M}_\alpha^{\text{RPax}}$:

$$\tilde{\mathbf{M}}_\alpha^{\text{RPax}} = \begin{pmatrix} {}^1\mathbf{M}_\alpha^{\text{RPax}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {}^3\mathbf{M}_\alpha^{\text{RPax}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & {}^3\mathbf{M}_\alpha^{\text{RPax}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & {}^3\mathbf{N}_\alpha^{\text{RPax}} \end{pmatrix} \quad (70)$$

with the expected spin-adapted blocks

$${}^1\mathbf{M}_\alpha^{\text{RPax}} = (\boldsymbol{\varepsilon} + \alpha{}^1\mathbf{A}' - \alpha{}^1\mathbf{B})^{1/2}(\boldsymbol{\varepsilon} + \alpha{}^1\mathbf{A}' + \alpha{}^1\mathbf{B}) \times (\boldsymbol{\varepsilon} + \alpha{}^1\mathbf{A}' - \alpha{}^1\mathbf{B})^{1/2}$$

and

$${}^3\mathbf{M}_\alpha^{\text{RPax}} = (\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' - \alpha{}^3\mathbf{B})^{1/2}(\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' + \alpha{}^3\mathbf{B}) \times (\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' - \alpha{}^3\mathbf{B})^{1/2}$$

along with the less expected last triplet block with opposite signs for ${}^3\mathbf{B}$

$${}^3\mathbf{N}_\alpha^{\text{RPax}} = (\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' + \alpha{}^3\mathbf{B})^{1/2}(\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' - \alpha{}^3\mathbf{B}) \times (\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' + \alpha{}^3\mathbf{B})^{1/2}$$

By eq 45, it gives the following spin-adaptation for the matrix $\mathbf{Q}_\alpha^{\text{RPax}}$

$$\tilde{\mathbf{Q}}_\alpha^{\text{RPax}} = \begin{pmatrix} {}^1\mathbf{Q}_\alpha^{\text{RPax}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & {}^3\mathbf{Q}_\alpha^{\text{RPax}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & {}^3\mathbf{Q}_\alpha^{\text{RPax}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & ({}^3\mathbf{Q}_\alpha^{\text{RPax}})^{-1} \end{pmatrix} \quad (71)$$

with the spin-adapted blocks ${}^1\mathbf{Q}_\alpha^{\text{RPax}} = (\boldsymbol{\varepsilon} + \alpha{}^1\mathbf{A}' - \alpha{}^1\mathbf{B})^{1/2}({}^1\mathbf{M}_\alpha^{\text{RPax}})^{-1/2}(\boldsymbol{\varepsilon} + \alpha{}^1\mathbf{A}' - \alpha{}^1\mathbf{B})^{1/2}$ and ${}^3\mathbf{Q}_\alpha^{\text{RPax}} = (\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' - \alpha{}^3\mathbf{B})^{1/2}({}^3\mathbf{M}_\alpha^{\text{RPax}})^{-1/2}(\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' - \alpha{}^3\mathbf{B})^{1/2}$. The last triplet block turns out to be the inverse $({}^3\mathbf{Q}_\alpha^{\text{RPax}})^{-1} = (\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' + \alpha{}^3\mathbf{B})^{1/2}({}^3\mathbf{N}_\alpha^{\text{RPax}})^{-1/2}(\boldsymbol{\varepsilon} + \alpha{}^3\mathbf{A}' + \alpha{}^3\mathbf{B})^{1/2}$ since according to eqs 25 and 35 one goes from \mathbf{Q}_α to \mathbf{Q}_α^{-1} by changing the sign of \mathbf{Y}_ω which is equivalent to changing the sign of \mathbf{B} .

The spin-adapted correlation energy expressions can be easily obtained by using the invariance of the trace under the transformation $\mathbf{C} \rightarrow \mathbf{U}^T\mathbf{C}\mathbf{U}$. The spin-adapted adiabatic-connection formula for the dRPA-I correlation energy is thus

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \int_0^1 d\alpha \operatorname{tr}\{[{}^1\mathbf{Q}_\alpha^{\text{dRPA}} - \mathbf{I}]^2\mathbf{K}\} \quad (72)$$

i.e., only singlet excitations contribute. Similarly, the corresponding plasmon formula contains only singlet excitation energies

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \sum_n ({}^1\omega_{1,n}^{\text{dRPA}} - {}^1\omega_n^{\text{dTDA}}) \quad (73)$$

The triplet term vanishes since both ${}^3\omega_{1,n}^{\text{dRPA}}$ and ${}^3\omega_n^{\text{dTDA}}$ are equal to the one-particle excitation energies $\varepsilon_a - \varepsilon_i$. Finally, the

spin-adapted alternative plasmon formula is

$$E_c^{\text{dRPA-I}} = \frac{1}{2} \text{tr}\{({}^1\mathbf{M}_1^{\text{dRPA}})^{1/2} - (\boldsymbol{\varepsilon} + {}^1\mathbf{K})\} \quad (74)$$

Both singlet and triplet excitations contribute the RPAX-II correlation energy. The spin-adapted adiabatic-connection formula for RPAX-II is

$$\begin{aligned} E_c^{\text{RPAX-II}} = & \frac{1}{4} \int_0^1 d\alpha \text{tr} \left\{ \frac{1}{2} ({}^1\mathbf{Q}_\alpha^{\text{RPAX}}) ({}^1\mathbf{A}' + {}^1\mathbf{B}) \right. \\ & \left. + \frac{1}{2} ({}^1\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} ({}^1\mathbf{A}' - {}^1\mathbf{B}) - {}^1\mathbf{A}' \right\} \\ & + \frac{3}{4} \int_0^1 d\alpha \text{tr} \left\{ \frac{1}{2} ({}^3\mathbf{Q}_\alpha^{\text{RPAX}}) ({}^3\mathbf{A}' + {}^3\mathbf{B}) \right. \\ & \left. + \frac{1}{2} ({}^3\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} ({}^3\mathbf{A}' - {}^3\mathbf{B}) - {}^3\mathbf{A}' \right\} \quad (75) \end{aligned}$$

The last triplet term gives a contribution identical to the other two triplet terms because the expression is invariant under the replacements $\mathbf{Q}_\alpha \rightarrow \mathbf{Q}_\alpha^{-1}$ and $\mathbf{B} \rightarrow -\mathbf{B}$. The spin-adaptation of the approximate RPAX-II correlation energy of eq 49 is

$$\begin{aligned} E_c^{\text{RPAX-IIa}} = & \frac{1}{4} \int_0^1 d\alpha \text{tr} \{ [{}^1\mathbf{Q}_\alpha^{\text{RPAX}} - \mathbf{I}] {}^1\mathbf{B} \} \\ & + \frac{2}{4} \int_0^1 d\alpha \text{tr} \{ [{}^3\mathbf{Q}_\alpha^{\text{RPAX}} - \mathbf{I}] {}^3\mathbf{B} \} \\ & - \frac{1}{4} \int_0^1 d\alpha \text{tr} \{ [({}^3\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} - \mathbf{I}] {}^3\mathbf{B} \} \quad (76) \end{aligned}$$

where now the last triplet term is not identical to the other two triplet terms. If we make the additional approximation $({}^3\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} \approx 2\mathbf{I} - {}^3\mathbf{Q}_\alpha^{\text{RPAX}}$, we arrive at the following expression:

$$\begin{aligned} E_c^{\text{RPAX-IIb}} = & \frac{1}{4} \int_0^1 d\alpha \text{tr} \{ [{}^1\mathbf{Q}_\alpha^{\text{RPAX}} - \mathbf{I}] {}^1\mathbf{B} \} \\ & + \frac{3}{4} \int_0^1 d\alpha \text{tr} \{ [{}^3\mathbf{Q}_\alpha^{\text{RPAX}} - \mathbf{I}] {}^3\mathbf{B} \} \quad (77) \end{aligned}$$

which could also have been obtained by starting from the spin-adapted formula of eq 75 and making the approximation $\mathbf{Q}_\alpha^{-1} \approx 2\mathbf{I} - \mathbf{Q}_\alpha$ in both the singlet and the triplet terms. The RPAX-II plasmon formula decomposes into sums over singlet and triplet excitation energies:

$$\begin{aligned} E_c^{\text{RPAX-II}} = & \frac{1}{4} \sum_n ({}^1\omega_{1,n}^{\text{RPAX}} - {}^1\omega_n^{\text{TDAX}}) \\ & + \frac{3}{4} \sum_n ({}^3\omega_{1,n}^{\text{RPAX}} - {}^3\omega_n^{\text{TDAX}}) \quad (78) \end{aligned}$$

and similarly for the alternative plasmon formula

$$\begin{aligned} E_c^{\text{RPAX-II}} = & \frac{1}{4} \text{tr}\{({}^1\mathbf{M}_1^{\text{RPAX}})^{1/2} - (\boldsymbol{\varepsilon} + {}^1\mathbf{A}')\} \\ & + \frac{3}{4} \text{tr}\{({}^3\mathbf{M}_1^{\text{RPAX}})^{1/2} - (\boldsymbol{\varepsilon} + {}^3\mathbf{A}')\} \quad (79) \end{aligned}$$

The last triplet term is identical to the other two because ${}^3\mathbf{N}_1^{\text{RPAX}}$ and ${}^3\mathbf{M}_1^{\text{RPAX}}$ have the same eigenvalues and thus $\text{tr}\{({}^3\mathbf{N}_1^{\text{RPAX}})^{1/2}\} = \text{tr}\{({}^3\mathbf{M}_1^{\text{RPAX}})^{1/2}\}$.

The spin-adapted dRPA-II correlation energy involves only singlet excitations:

$$\begin{aligned} E_c^{\text{dRPA-II}} = & \frac{1}{2} \int_0^1 d\alpha \text{tr} \left\{ \frac{1}{2} ({}^1\mathbf{Q}_\alpha^{\text{dRPA}}) ({}^1\mathbf{A}' + {}^1\mathbf{B}) \right. \\ & \left. + \frac{1}{2} ({}^1\mathbf{Q}_\alpha^{\text{dRPA}})^{-1} ({}^1\mathbf{A}' - {}^1\mathbf{B}) - {}^1\mathbf{A}' \right\} \quad (80) \end{aligned}$$

since for the triplet blocks ${}^3\mathbf{Q}_\alpha^{\text{dRPA}} = \mathbf{I}$ and the contribution vanishes. Likewise, the spin-adaptation of the approximate dRPA-II correlation energy of eq 59 is simply

$$E_c^{\text{dRPA-IIa}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \{ [{}^1\mathbf{Q}_\alpha^{\text{dRPA}} - \mathbf{I}] {}^1\mathbf{B} \} \quad (81)$$

Finally, the spin-adapted RPAX-I correlation energy expression is

$$E_c^{\text{RPAX-I}} = \frac{1}{2} \int_0^1 d\alpha \text{tr} \{ [{}^1\mathbf{Q}_\alpha^{\text{RPAX}} - \mathbf{I}] {}^1\mathbf{K} \} \quad (82)$$

where only singlet excitations contribute since the triplet blocks of the matrix \mathbf{K} are zero.

6. NUMERICAL ILLUSTRATIONS

The above-described spin-adapted RPA correlation energy variants based on the adiabatic-connection formula have been implemented in the development version of the MOLPRO quantum chemistry package.⁵³ The numerical equality of the alternative but equivalent expressions has been carefully tested and has been confirmed within the usual accuracy of quantum chemical calculations. In each case, we start by doing a usual Kohn–Sham (KS) calculation with some approximate density functional and evaluate the RPA correlation energy with the KS orbitals. The total RPA energy is calculated as

$$E_{\text{tot}}^{\text{RPA}} = E_{\text{EXX}} + E_c^{\text{RPA}} \quad (83)$$

where E_{EXX} is the exact exchange (EXX) energy expression evaluated with the same KS orbitals. This exchange energy is Hartree–Fock type, and it is not to be confused with the optimized effective potential (OEP) type local exchange, often denoted by the same acronym. For comparison, we also perform range-separated calculations, in which we start from a range-separated hybrid (RSH),⁴⁵ using the short-range PBE exchange-correlation functional of ref 54, and add the long-range RPA correlation energy evaluated with RSH orbitals

$$E_{\text{tot}}^{\text{RSH+RPA}} = E_{\text{RSH}} + E_c^{\text{lr,RPA}} \quad (84)$$

The long-range RPA correlation energy $E_c^{\text{lr,RPA}}$ is calculated by replacing the Coulombic two-electron integrals by the two-electron integrals with the long-range interaction $\text{erf}(\mu r)/r$, just as in refs 15, 31, and 33. We use a fixed value of the range-separation parameter of $\mu = 0.5 \text{ bohr}^{-1}$. This value corresponds to a reasonable global compromise, as was shown previously⁵⁵ in a study of thermochemical properties, and as was confirmed later by using alternative criteria leading to similar estimates of the μ parameter (see, e.g., ref 56). In all cases, the adiabatic-connection integration is performed by an eight-point Gauss–Legendre quadrature.

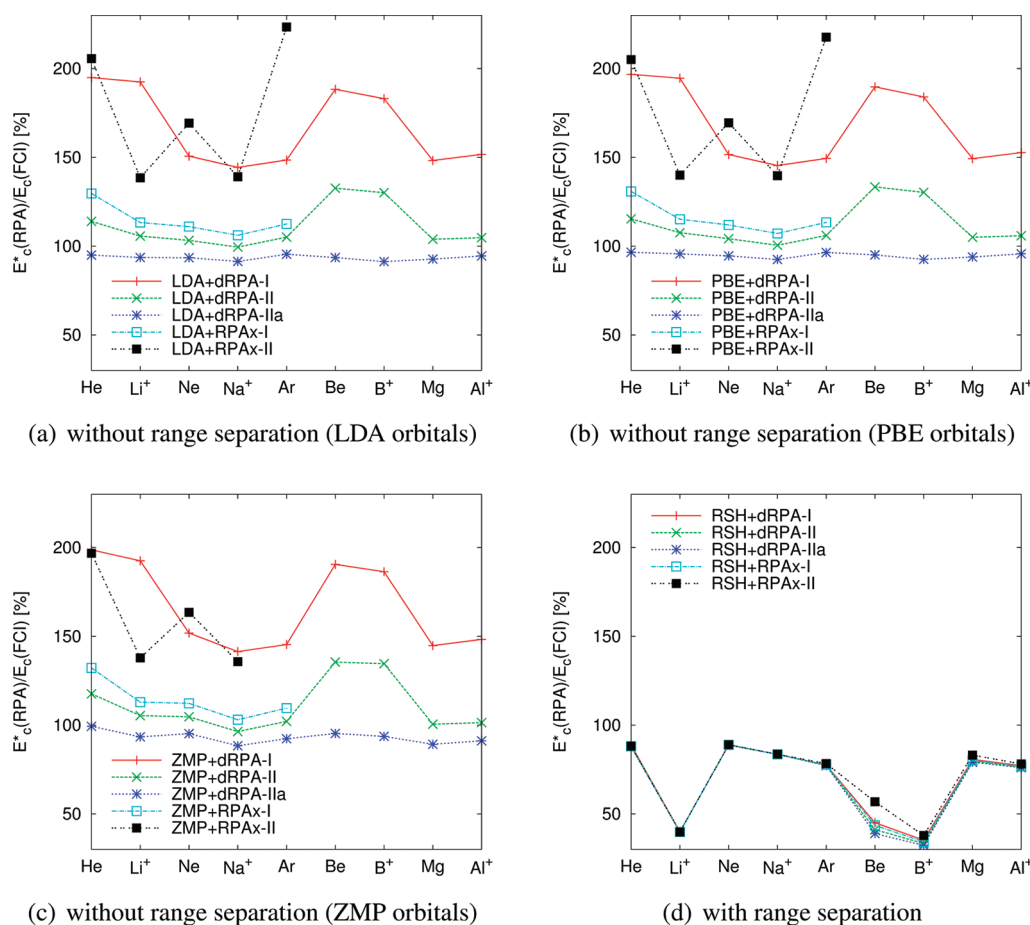


Figure 1. Ratios between various RPA correlation energy variants and the FCI-quality correlation energy as estimated by Davidson and co-workers,^{60,61} with and without range separation. All of the correlation energies have been extrapolated to the CBS limit. The RPA correlation energies $E^*_c(\text{RPA})$ are redefined here as the difference between the total RPA energies and the regular HF energies.

The RPA correlation energies are extrapolated to the complete basis set (CBS) limit by the usual $1/X^3$ formula⁵⁷ for a series of Dunning basis sets. In contrast to the usual two-point extrapolation procedure,^{58,59} all of the available correlation energies calculated by at least a triple- ζ basis set are used. The single-determinant reference energies are evaluated with a large basis set so that they can be considered converged.

6.1. Atomic Correlation Energies. As a first test, we have calculated correlation energies for a series of atoms and atomic cations and compared them with full configuration interaction (FCI) quality correlation energies as estimated by Davidson and co-workers.^{60,61} In order to make a direct comparison with the FCI-quality correlation energies which are defined with respect to the HF energies, we redefine RPA correlation energies as the difference between the total RPA energies and the regular HF energies. The single-determinant reference energies are calculated with a large even-tempered basis set. With this basis set, the HF energies agree within all significant digits with Davidson's reference data. Core excitations are included in the calculation of the RPA correlation energies and are extrapolated from the series of aug-cc-pCVXZ basis sets for He up to $X = 6$; for B^+ , Al^+ , Ne, and Ar, up to $X = 5$; and for Li^+ , Na^+ , Be, and Mg, up to $X = Q$.

Figure 1a–c show the ratios of the correlation energies for each full-range RPA variant (dRPA-I, dRPA-II, dRPA-IIa, RPAX-I, RPAX-II) to the FCI-quality correlation energies, using orbitals

obtained with the local-density approximation (LDA),⁶² the Perdew–Burke–Ernzerhof (PBE),⁶³ and the Zhao–Morrison–Parr (ZMP)⁶⁴ exchange–correlation potentials. The ZMP potentials have been constructed from high-quality *ab initio* wave functions (Brueckner coupled cluster doubles).⁶⁵ It appears that the correlation energies are only marginally dependent on the quality of the KS orbitals, at least for this series of atomic systems. The full-range RPAX-I and RPAX-II variants suffer from instabilities in the RPAX response equation for the Be, B^+ , Mg, and Al^+ systems, and additionally Ar in the case of RPAX-II with the ZMP orbitals. In fact, the strongly overestimated RPAX-II correlation energies of Ar obtained with the LDA and PBE orbitals indicate a situation close to instability. More generally, the presence of near instabilities may be considered as being at the origin of the relatively strong overestimation (usually more than 150%) of the correlation energy in RPAX-II. In view of the poor performance of RPAX-II, we did not test the approximate versions of eqs 76 and 77. The RPAX-I variant only involves singlet excitations and thus is not subject to triplet instabilities. It gives quite reasonable correlation energies (maximum 25% of overestimation) for He, Li^+ , Ne, Na^+ , and even Ar. However, RPAX-I is subject to singlet instabilities, which appear for the rest of the systems. The dRPA-I variant is free of any instability problems, since the dRPA response matrix is positive definite by construction, but has nevertheless a tendency for overestimating correlation energies

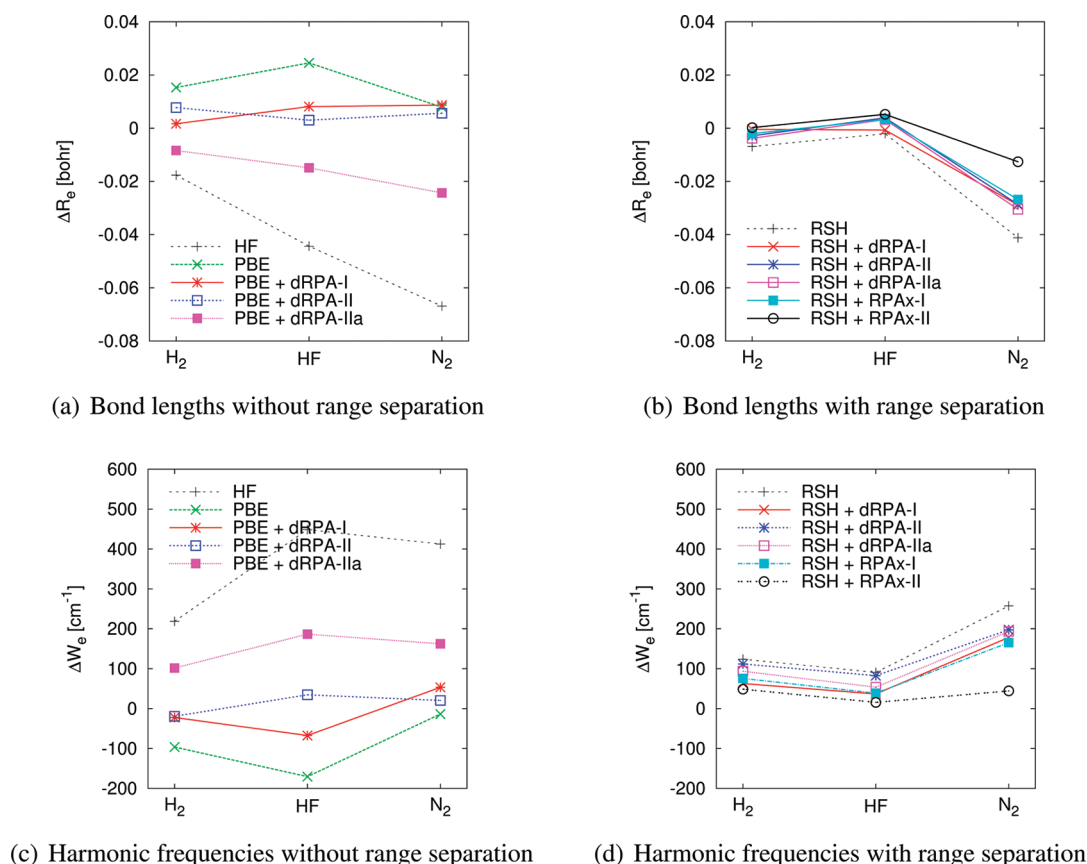


Figure 2. Errors in the equilibrium bond lengths and harmonic vibrational frequencies for simple diatomic molecules, calculated with the full-range and range-separated RPA variants and compared to experimental reference values. All of the correlation energies have been extrapolated to the CBS limit. The experimental reference values are (in bohr and cm^{-1}) H_2 , $R_e = 1.40112$, $\omega_e = 4401.21$; HF, $R_e = 1.73250$, $\omega_e = 4138.32$; N_2 , $R_e = 2.07431$, $\omega_e = 2358.57$.⁶⁶

by a factor of 1.5 to 2. This systematic error can be easily corrected by including exchange in the energy expression. In fact, the dRPA-II variant and especially its approximate form dRPA-IIa (AC-SOSEX) lead to a very good reproduction of the reference correlation energies. Similar effects could be observed recently in the direct ring-CCD (dRPA-I) and SOSEX calculations of correlation energies of Klopper et al.,⁴⁰ performed with a much smaller basis set.

As mentioned previously, dRPA-IIa (or AC-SOSEX) and the ring-CCD-based SOSEX correlation energies are expected to be quite close to each other. Numerical results (not shown in the figures) confirm this expectation. For two-electron systems (He, Li^+), the dRPA-IIa and SOSEX correlation energies are identical, while for the rest of the systems, the relative difference is less than 0.15%. The largest absolute difference, 0.82 mHartree, has been found in full-range calculations on the Ar atom. It is interesting to note that the ring-CCD-based SOSEX correlation energies are always deeper than the dRPA-IIa values. This fact cannot be interpreted simply by the comparison of the third order energy expressions, reported in ref 34.

Figure 1d shows the same total correlation energies obtained with range separation, i.e., the sum of the short-range PBE correlation energy and the long-range RPA correlation energy. The situation is quite different from the full-range RPA calculations. First, we do not encounter any instability problems anymore. Second, all of the range-separated RPA variants give essentially identical correlation energies. Third, the correlation

energies are systematically underestimated, for most of the systems with less than 20% of error, but with the notable exceptions of Li^+ , Be, and B^+ , for which the correlation energies are underestimated by as much as 50%. These findings may be due to the fact that the systems considered here have very compact densities, and for the value of the range separation used here, $\mu = 0.5 \text{ bohr}^{-1}$, the major part of correlation is assigned to the short-range density functional rather than to the long-range RPA calculation. Improvement over these results would require either increasing the value of μ or using a more accurate short-range density-functional approximation.

6.2. Bond Lengths and Harmonic Vibrational Frequencies.

Figure 2 reports equilibrium bond lengths and harmonic vibrational frequencies calculated with the full-range and range-separated RPA variants for three simple diatomic molecules, representing an apolar single bond (H_2), a strongly polar single bond (HF), and an apolar multiple bond (N_2). The full-range RPA calculations are done with PBE orbitals, while the range-separated RPA calculations are done with the short-range PBE density functional. All RPA calculations are without core excitations and extrapolated to the CBS limit with the series of basis sets aug-cc-pVXZ with $X = \text{T}, \text{Q},$ and 5 . The single-determinant reference energies are calculated with the aug-cc-pVSZ basis set. Due to instabilities in the full-range RPAX response equation, only the full-range dRPA values can be calculated, while no instabilities are found for the range-separated RPAX calculations. Without range separation, big differences are found between the

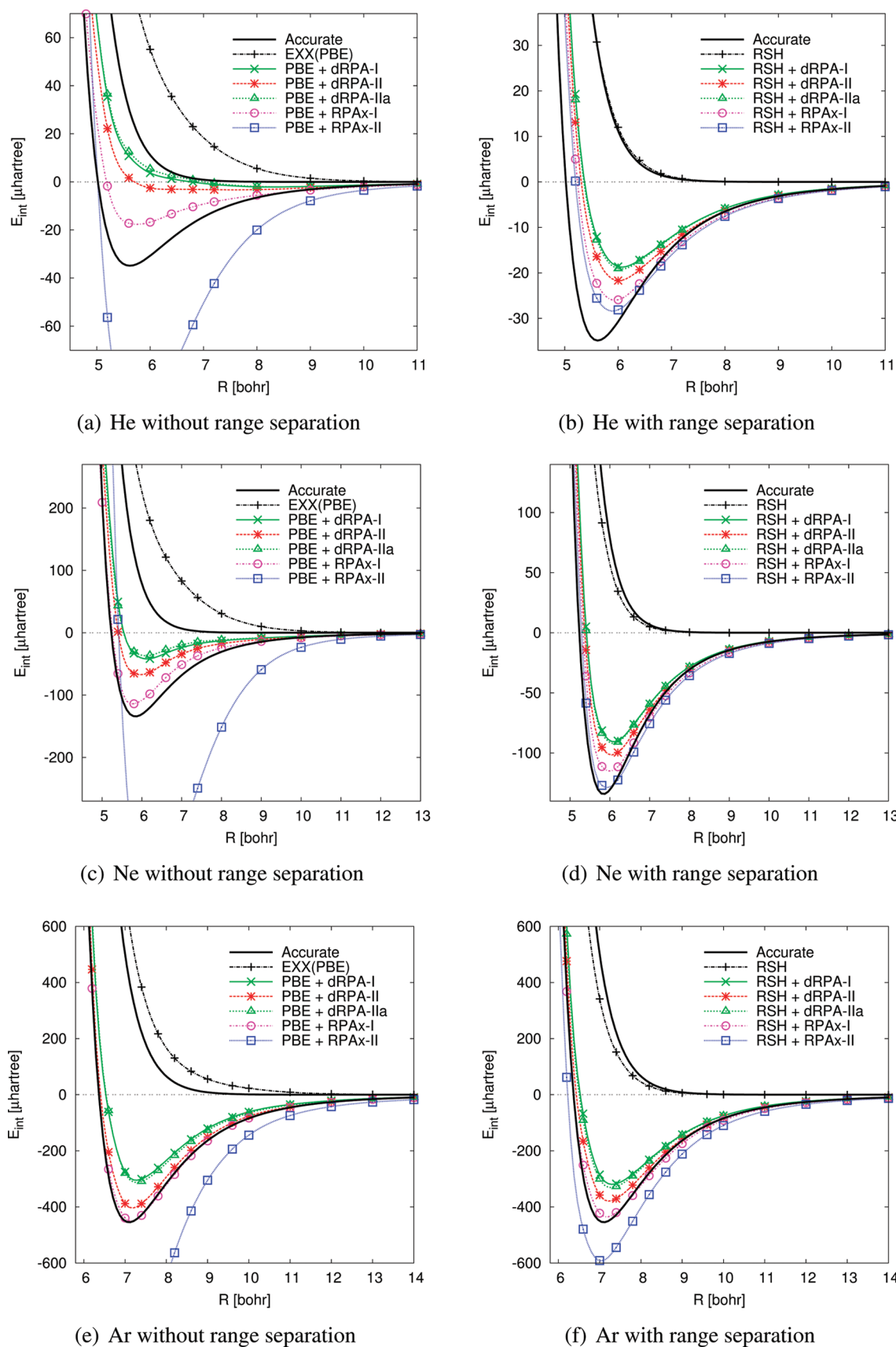


Figure 3. Interaction energy curves of He_2 , Ne_2 , and Ar_2 , calculated with the full-range and range-separated RPA variants. All of the correlation energies have been extrapolated to the CBS limit.

different methods. The dRPA-I and dRPA-II variants perform quite well and represent an important improvement over both HF and KS PBE. The approximate variant dRPA-IIa is significantly less accurate than dRPA-II. With range separation, the methods give much closer results to one another. The best range-separated variant for this small set of bond lengths and harmonic frequencies appears to be RPax-II, especially in the case of the N_2 molecule.

6.3. London Dispersion Interactions. Figure 3 shows the interaction energy curves of the three rare-gas dimers He_2 , Ne_2 , and Ar_2 , calculated with the full-range and range-separated RPA variants. The full-range RPA calculations are done with PBE orbitals, while the range-separated RPA calculations are done with the short-range PBE density functional. All RPA calculations are without core excitations and extrapolated to the CBS limit with the series of basis sets aug-cc-pVXZ with $X = T, Q, S,$ and 6 . The single-determinant reference energies are obtained with the aug-cc-pV6Z basis set. We note that when using LDA orbitals (not shown), instabilities are found for Ne_2 and Ar_2 in a wide range of interatomic distances. In contrast, no instabilities are encountered in the case of PBE, neither with nor without range separation.

The continuous curves without points represent, on the one hand, the accurate reference curves according to the analytical potential energy expression of Tang and Toennies⁶⁷ and, on the other hand, the repulsive (exponentially decaying) component of the same potential. These latter curves serve as useful guides to estimate the accuracy of the single-determinant reference energies, i.e., EXX energies with PBE orbitals or RSH energies. It is quite clear that the quality of the results depends strongly on the quality of the repulsive curve. The poorest interaction energy curves are obtained for the He_2 dimer without range separation, for which the EXX energy is too strongly repulsive. The prerequisite of the good performance of the range-separated calculations is obviously the excellent accuracy of the RSH energy, which, for He_2 , is in almost perfect agreement with the reference repulsive curve.

The full-range RPax-II variant overestimates systematically the binding energy by a factor of 3 or more. The dRPA-I method largely underestimates the interaction energies, and for He_2 , it does not provide any minimum at all, although the non-binding character is mostly due to the bad single-determinant energy. The dRPA-II variant systematically gives more binding than dRPA-I but also tends to underestimate the interaction energies. The approximate dRPA-IIa variant gives results that are always very close to those of dRPA-I. This is not surprising since the dRPA-I and the dRPA-IIa methods differ only by the presence of exponentially decaying exchange integrals in the interaction matrix which become quite rapidly negligible to the interaction energy in van der Waals complexes. This behavior is analogous to that of the SOSEX method, which gives dispersion interaction energies also very close to those of dRPA-I.⁴⁶ The best full-range method for these rare gas dimers is RPax-I, which is in quite good agreement with the reference curves for Ne_2 and Ar_2 .

With range separation, all of the RPA variants give much closer interaction energy curves to each other, but the same trends are found. Range-separated dRPA-I, dRPA-II, and dRPA-IIa methods systematically underestimate interaction energies. The range-separated RPax-II significantly overbinds Ar_2 , and the range-separated RPax-I globally gives the most accurate interaction energies.

7. CONCLUSIONS

We have analyzed several RPA correlation energy variants based on the adiabatic-connection formula: dRPA-I, dRPA-II, RPax-I, and RPax-II. These variants have the generic form of an interaction-strength-averaged two-particle density matrix contracted with two-electron integrals. They differ in the way the exchange interactions are treated. The dRPA-I variant is just the usual RPA of the density-functional/material-science community and neglects all exchange interactions. The dRPA-II variant uses a density matrix without exchange but contracted with antisymmetrized two-electron integrals. It is original to this work, although it resembles the SOSEX method,²³ especially in its approximate form named dRPA-IIa. The RPax-I uses a density matrix with exchange but contracted with nonantisymmetrized two-electron integrals. It has previously been discussed in the context of range-separated density-functional theory.^{15,33} The RPax-II variant uses a density matrix with exchange and contracted with antisymmetrized two-electron integrals. The RPax-II method itself is obviously not new,⁴⁹ but we have derived several new expressions for it. Contracting the density matrix with either nonantisymmetrized or antisymmetrized two-electron integrals is not equivalent because of the breaking of the antisymmetry of the density matrix in RPA. For the dRPA-I and RPax-II variants, we have made the connection with the plasmon formulation and clarify the origin of the factor of 1/4 in the plasmon formula for RPax-II instead of the factor of 1/2 for dRPA-I. We have carefully studied the second-order limit in the electron–electron interaction and shown that all of the correlation energy variants except for dRPA-I correctly reduce to the MP2 correlation energy (see the Appendix). Finally, we have derived the spin-adapted forms of all of these methods for closed-shell systems and implemented and tested them with and without range separation of the electron–electron interaction.

The numerical examples on atomic and molecular systems show that the RPax variants without range separation frequently suffer from instabilities in the RPax response equation, which make it impossible to extract a meaningful correlation energy in these cases. However, no instabilities are encountered with range separation, and the RPax variants can be thus safely applied. The tests performed do not allow us to identify an RPA variant which would be uniformly better than the others. Without range-separation, dRPA-II performs well for atomic correlation energies and equilibrium molecular properties but significantly underestimates London dispersion interaction energies for which RPax-I is more accurate. With range separation, all of the RPA variants tend to give more accurate results, and they also become much more similar to each other. Range-separated RPax-II appears as the best variant for equilibrium molecular properties, and range-separated RPax-I is the best variant for dispersion interaction energies.

We hope that the overview of the RPA correlation energy variants provided in this work will be useful for a better understanding of RPA methods and can serve as a starting point for the design of improved approximations.

APPENDIX

Appendix. Second-Order Approximations to the RPA Correlation Energy Expressions

In this appendix, we explicitly derive the approximations at second order in the electron–electron interaction of the RPA correlation energy variants.

We will deal with the more general RPax response equation and obtain dRPA as a special case. We thus start from the response equation:

$$(\Lambda_0 + \alpha \mathbb{W})C_{\alpha,n} = \omega_{\alpha,n} \Delta C_{\alpha,n} \quad (85)$$

with

$$\Lambda_0 = \begin{pmatrix} \boldsymbol{\varepsilon} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varepsilon} \end{pmatrix}, \mathbb{W} = \begin{pmatrix} \mathbf{A}' & \mathbf{B} \\ \mathbf{B} & \mathbf{A}' \end{pmatrix}, \Delta = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \quad (86)$$

where $\boldsymbol{\varepsilon}$ is a diagonal matrix composed of orbital energy differences $\varepsilon_{ia} = \varepsilon_a - \varepsilon_i$ and \mathbf{A}' and \mathbf{B} are matrices composed of the the antisymmetrized two-electron integrals $A'_{ia,jb} = \langle ib || aj \rangle$ and $B_{ia,jb} = \langle ab || ij \rangle$, and \mathbf{I} is the identity matrix. We assume that all occupied (denoted by i and j) and all virtual (a and b) orbitals are real. In the following, the index pairs ia and jb will be replaced with simple indices m and n . Note that the matrices are symmetric: $A'_{n,m} = A'_{m,n}$ and $B_{n,m} = B_{m,n}$. The solutions of eq 85 come in pairs; i.e., if $C_{\alpha,n} = (\mathbf{x}_{\alpha,n}, \mathbf{y}_{\alpha,n})$ is an eigenvector with a positive eigenvalue $\omega_{\alpha,n} > 0$, then $C_{\alpha,-n} = (\mathbf{y}_{\alpha,n}, \mathbf{x}_{\alpha,n})$ is an eigenvector with the negative eigenvalue $\omega_{\alpha,-n} = -\omega_{\alpha,n}$. In the following, we will use positive integer indices to denote solutions which connect to positive eigenvalues in the limit of a vanishing coupling parameter α , i.e., to $\omega_{0,n} > 0$. Note that we also suppose a nonvanishing HOMO–LUMO gap.

The positive energy solutions of eq 85 for $\alpha = 0$ are trivially given by $\omega_{0,n} = \varepsilon_m$, $\mathbf{x}_{0,n} = \mathbf{1}_m$ and $\mathbf{y}_{0,n} = \mathbf{0}$, where $\mathbf{1}_n$ denotes the n th unit vector, i.e., a vector with vanishing components except for the n th component which is equal to 1. We now wish to find the first-order correction $C_n^{(1)}$ to the eigenvector employing the power-series Ansatz

$$\omega_{\alpha,n} = \omega_{0,n} + \alpha \omega_n^{(1)} + \dots \quad (87)$$

$$C_{\alpha,n} = C_{0,n} + \alpha C_n^{(1)} + \dots \quad (88)$$

Plugging this into eq 85, one sees that the first-order corrections are obtained from solving

$$\Lambda_0 C_n^{(1)} + \mathbb{W} C_{0,n} = \omega_{0,n} \Delta C_n^{(1)} + \omega_n^{(1)} \Delta C_{0,n} \quad (89)$$

Multiplication of this equation from the left with $C_{0,n}^T$ and using $C_{0,n}^T \Lambda_0 C_n^{(1)} = \omega_{0,n} C_{0,n}^T \Delta C_n^{(1)}$ along with the normalization condition $C_{0,n}^T \Delta C_{0,n} = 1$ gives the first-order correction to the eigenvalue

$$\omega_n^{(1)} = C_{0,n}^T \mathbb{W} C_{0,n} = A'_{n,n} \quad (90)$$

Multiplying eq 89 from the left with $C_{0,m}^T$ for $m \neq n$, using $C_{0,m}^T \Lambda_0 C_n^{(1)} = \omega_{0,m} C_{0,m}^T \Delta C_n^{(1)}$, and employing the orthogonality condition $C_{0,m}^T \Delta C_{0,n} = 0$ leads to

$$C_{0,m}^T \Delta C_n^{(1)} = -\frac{C_{0,m}^T \mathbb{W} C_{0,n}}{\omega_{0,m} - \omega_{0,n}} \quad (91)$$

provided that the zeroth-order eigenvalues are non-degenerate, i.e., that no two occupied–virtual orbital energy differences match. Repeating the same operations for $C_{0,-m}^T$ one arrives at

$$C_{0,-m}^T \Delta C_n^{(1)} = \frac{C_{0,-m}^T \mathbb{W} C_{0,n}}{\omega_{0,m} + \omega_{0,n}} \quad (92)$$

where $\omega_{0,-m} = -\omega_{0,m}$ has been used. Using the resolution of identity, $1 = \sum_m C_{0,m} C_{0,m}^T + C_{0,-m} C_{0,-m}^T$ the orthogonality of $C_n^{(1)}$ to the zeroth-order eigenvector, i.e., $C_{0,n}^T \Delta C_n^{(1)} = 0$, and $\Delta^2 = 1$, we find the expansion of the first-order correction to the positive-energy eigenvectors

$$C_n^{(1)} = -\sum_{m \neq n} \frac{C_{0,m}^T \mathbb{W} C_{0,n}}{\omega_{0,m} - \omega_{0,n}} \Delta C_{0,m} + \sum_m \frac{C_{0,-m}^T \mathbb{W} C_{0,n}}{\omega_{0,m} + \omega_{0,n}} \Delta C_{0,-m} \quad (93)$$

From eq 93, it follows that the first-order corrections read more explicitly

$$\mathbf{x}_n^{(1)} = -\sum_{m \neq n} \frac{A'_{m,n}}{\varepsilon_m - \varepsilon_n} \mathbf{1}_m \quad (94a)$$

$$\mathbf{y}_n^{(1)} = -\sum_m \frac{B_{m,n}}{\varepsilon_m + \varepsilon_n} \mathbf{1}_m \quad (94b)$$

The first-order corrections to the negative-energy solutions are simply $\omega_{-n}^{(1)} = -\omega_n^{(1)}$, $\mathbf{x}_{-n}^{(1)} = \mathbf{y}_n^{(1)}$, and $\mathbf{y}_{-n}^{(1)} = \mathbf{x}_n^{(1)}$.

We can obtain the first-order expansion of the matrix $\mathbf{Q}_\alpha^{\text{RPax}}$

$$\begin{aligned} \mathbf{Q}_\alpha^{\text{RPax}} &= \sum_n (\mathbf{x}_{\alpha,n} + \mathbf{y}_{\alpha,n})(\mathbf{x}_{\alpha,n} + \mathbf{y}_{\alpha,n})^T \\ &= \sum_n \mathbf{1}_n \mathbf{1}_n^T + \alpha \sum_n [\mathbf{x}_n^{(1)} \mathbf{1}_n^T + \mathbf{1}_n \mathbf{x}_n^{(1)T} + \mathbf{y}_n^{(1)} \mathbf{1}_n^T \\ &\quad + \mathbf{1}_n \mathbf{y}_n^{(1)T}] + O(\alpha^2) \end{aligned} \quad (95)$$

where the sum over n refers to positive-energy eigenvectors only. The first term is simply the identity matrix

$$\sum_n \mathbf{1}_n \mathbf{1}_n^T = \mathbf{I} \quad (96)$$

Using eq 94a, one can show that the term depending on $\mathbf{x}_n^{(1)}$ vanishes:

$$\begin{aligned} \sum_n \mathbf{x}_n^{(1)} \mathbf{1}_n^T + \mathbf{1}_n \mathbf{x}_n^{(1)T} &= -\sum_n \sum_{m \neq n} \frac{A'_{m,n}}{\varepsilon_m - \varepsilon_n} \mathbf{1}_m \mathbf{1}_n^T \\ &\quad - \sum_n \sum_{m \neq n} \frac{A'_{m,n}}{\varepsilon_m - \varepsilon_n} \mathbf{1}_n \mathbf{1}_m^T = 0 \end{aligned} \quad (97)$$

This is seen by swapping n and m in the last term and noting that $A'_{m,n}/(\varepsilon_m - \varepsilon_n)$ is antisymmetric when exchanging m and n . Finally, using eq 94b, the term depending on $\mathbf{y}_n^{(1)}$ gives

$$\begin{aligned} \sum_n \mathbf{y}_n^{(1)} \mathbf{1}_n^T + \mathbf{1}_n \mathbf{y}_n^{(1)T} &= -\sum_n \sum_m \frac{B_{m,n}}{\varepsilon_m + \varepsilon_n} \mathbf{1}_m \mathbf{1}_n^T \\ &\quad - \sum_n \sum_m \frac{B_{m,n}}{\varepsilon_m + \varepsilon_n} \mathbf{1}_n \mathbf{1}_m^T = -2\bar{\mathbf{B}}, \end{aligned} \quad (98)$$

where $\bar{\mathbf{B}}$ is the matrix with elements $\bar{B}_{m,n} = B_{m,n}/(\varepsilon_m + \varepsilon_n)$ or, more explicitly, $\bar{B}_{ia,jb} = B_{ia,jb}/(\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j)$. Therefore, we have

$$\mathbf{Q}_\alpha^{\text{RPax}} = \mathbf{I} - 2\alpha \bar{\mathbf{B}} + O(\alpha^2) \quad (99)$$

and, similarly, the first-order expansion of the inverse matrix $(\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} = \sum_n (\mathbf{x}_{\alpha,n} - \mathbf{y}_{\alpha,n})(\mathbf{x}_{\alpha,n} - \mathbf{y}_{\alpha,n})^T$ yields

$$(\mathbf{Q}_\alpha^{\text{RPAX}})^{-1} = \mathbf{I} + 2\alpha\bar{\mathbf{B}} + O(\alpha^2) \quad (100)$$

Equations 99 and 100 show that the approximation $\mathbf{Q}_\alpha + \mathbf{Q}_\alpha^{-1} \approx 2\mathbf{I}$, which leads to the definitions of $E_c^{\text{RPAX-IIa}}$ [eq 49] and $E_c^{\text{dRPA-IIa}}$ [eq 59], is correct up to first order in α .

All of the above considerations remain valid for the dRPA case, except for the replacements $\mathbf{A}' \rightarrow \mathbf{K}$ and $\mathbf{B} \rightarrow \mathbf{K}$, with the obvious results

$$\mathbf{Q}_\alpha^{\text{dRPA}} = \mathbf{I} - 2\alpha\bar{\mathbf{K}} + O(\alpha^2) \quad (101)$$

and

$$(\mathbf{Q}_\alpha^{\text{dRPA}})^{-1} = \mathbf{I} + 2\alpha\bar{\mathbf{K}} + O(\alpha^2) \quad (102)$$

where the matrix elements of $\bar{\mathbf{K}}$ are given by $\bar{K}_{m,n} = K_{m,n}/(\varepsilon_m + \varepsilon_n)$ or, more explicitly, $\bar{K}_{ia,jb} = K_{ia,jb}/(\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j)$.

We can give now the second-order limits of the RPA correlation energy variants. Using eq 101, we find the second-order limit of the dRPA correlation energy variant of eq 28:

$$E_c^{\text{dRPA-I}} \approx \frac{1}{2} \int_0^1 d\alpha \text{tr}\{[-2\alpha\bar{\mathbf{K}}]\mathbf{K}\} = -\frac{1}{2} \text{tr}\{\bar{\mathbf{K}}\mathbf{K}\} \quad (103)$$

which is not the normal MP2 correlation energy but a MP2-like correlation energy without exchange, also called direct MP2 or JMP2.⁵⁰ In a similar way, eqs 101 and 100 give the second-order limit of the RPAX-II correlation energy variant of eq 48, which is the same for its approximation of eq 49:

$$E_c^{\text{RPAX-II}} \approx E_c^{\text{RPAX-IIa}} \approx \frac{1}{4} \int_0^1 d\alpha \text{tr}\{[-2\alpha\bar{\mathbf{B}}]\mathbf{B}\} = -\frac{1}{4} \text{tr}\{\bar{\mathbf{B}}\mathbf{B}\} \quad (104)$$

which is exactly the MP2 correlation energy expression (except for the possible replacement of Hartree–Fock orbitals and orbital energies with corresponding Kohn–Sham quantities). The second-order limit of the dRPA-II correlation energy variant of eq 58 and its approximation of eq 59 are found with eqs 101 and 102:

$$E_c^{\text{dRPA-II}} \approx E_c^{\text{dRPA-IIa}} \approx \frac{1}{2} \int_0^1 d\alpha \text{tr}\{[-2\alpha\bar{\mathbf{K}}]\mathbf{B}\} = -\frac{1}{2} \text{tr}\{\bar{\mathbf{K}}\mathbf{B}\} \quad (105)$$

Using the antisymmetry of \mathbf{B} and observing the prefactor of 1/2, it can easily be seen that this is another way to write the usual MP2 correlation energy expression. Finally, the RPAX-I correlation energy variant of eq 61 has the following second-order limit:

$$E_c^{\text{RPAX-I}} \approx \frac{1}{2} \int_0^1 d\alpha \text{tr}\{[-2\alpha\bar{\mathbf{B}}]\mathbf{K}\} = -\frac{1}{2} \text{tr}\{\bar{\mathbf{B}}\mathbf{K}\} \quad (106)$$

which again exactly corresponds to the usual MP2 correlation energy expression.

Let us now consider the case of a closed-shell system. In this case, there is (at least) a 4-fold degeneracy in the ε block of Λ_0 since $\varepsilon_{it} = \varepsilon_{it}$ and $\varepsilon_{at} = \varepsilon_{at}$. As a consequence, the condition of nondegeneracy of zeroth-order excitation energies $\omega_{0,n} = \varepsilon_{ia}$ leading to eqs 91 and 94a is violated. Even if the final results for the second-order correlation energies do not contain differences of excitation energies anymore, a different derivation is needed. This may be achieved by first spin-adapting the RPA response

equation (for the details, see, e.g., ref 33) and only subsequently making the perturbation expansion on the spin-adapted energy expressions of section 5. Assuming the absence of further degeneracies between orbital energy differences (zeroth-order excitation energies), one obtains formally identical expansions for the singlet and triplet blocks. For example, the spin-adapted matrices ${}^1\mathbf{Q}_\alpha = \sum_n ({}^1\mathbf{x}_{\alpha,n} + {}^1\mathbf{y}_{\alpha,n})({}^1\mathbf{x}_{\alpha,n} + {}^1\mathbf{y}_{\alpha,n})^T$ and ${}^3\mathbf{Q}_\alpha = \sum_n ({}^3\mathbf{x}_{\alpha,n} + {}^3\mathbf{y}_{\alpha,n})({}^3\mathbf{x}_{\alpha,n} + {}^3\mathbf{y}_{\alpha,n})^T$, where $({}^1\mathbf{x}_{\alpha,n}, {}^1\mathbf{y}_{\alpha,n})$ and $({}^3\mathbf{x}_{\alpha,n}, {}^3\mathbf{y}_{\alpha,n})$ are the singlet and triplet eigenvectors, and the corresponding inverse matrices $({}^1\mathbf{Q}_\alpha)^{-1}$ and $({}^3\mathbf{Q}_\alpha)^{-1}$ have the following expansions in the case of RPAX:

$$({}^{1,3}\mathbf{Q}_\alpha^{\text{RPAX}})^{\pm 1} = \mathbf{I} \mp 2\alpha{}^{1,3}\bar{\mathbf{B}} + O(\alpha^2) \quad (107)$$

with ${}^1\bar{\mathbf{B}}_{m,n} = {}^1\mathbf{B}_{m,n}/(\varepsilon_m + \varepsilon_n)$ and ${}^3\bar{\mathbf{B}}_{m,n} = {}^3\mathbf{B}_{m,n}/(\varepsilon_m + \varepsilon_n)$. Using these results, one can easily check that all of the spin-adapted correlation expressions of section 5 correctly reduce to MP2 at second order, except for the dRPA-I variant, which reduces to direct MP2.

AUTHOR INFORMATION

Corresponding Author

*E-mail: angyan@crm2.uhp-nancy.fr; julien.toulouse@upmc.fr; georg.jansen@uni-due.de.

Present Addresses

^{||}Department of Chemistry and Biochemistry, M/C 9510, University of California, Santa Barbara, California 93106, United States

ACKNOWLEDGMENT

We thank A. Savin (Paris) for discussions. This work was supported by ANR (Agence National de Recherche) via contract number ANR-07-BLAN-0272 (Wademecom). G.J. thanks Nancy University for a visiting professorship during which part of this work was carried out.

REFERENCES

- (1) Yan, Z.; Perdew, J. P.; Kurth, S. *Phys. Rev. B* **2000**, *61*, 16430–16439.
- (2) Furche, F. *Phys. Rev. B* **2001**, *64*, 195120.
- (3) Aryasetiawan, F.; Miyake, T.; Terakura, K. *Phys. Rev. Lett.* **2002**, *88*, 166401.
- (4) Miyake, T.; Aryasetiawan, F.; Kotani, T.; Schilfgaarde, M. v.; Usuda, M.; Terakura, K. *Phys. Rev. B* **2002**, *66*, 245103.
- (5) Fuchs, M.; Gonze, X. *Phys. Rev. B* **2002**, *65*, 235109.
- (6) Niquet, Y. M.; Gonze, X. *Phys. Rev. B* **2004**, *70*, 245115.
- (7) Fuchs, M.; Niquet, Y. M.; Gonze, X.; Burke, K. *J. Chem. Phys.* **2005**, *122*, 094116.
- (8) Furche, F.; van Voorhis, T. *J. Chem. Phys.* **2005**, *122*, 164106.
- (9) Dahlen, N. E.; van Leeuwen, R.; von Barth, U. *Phys. Rev. A* **2006**, *73*, 012511.
- (10) Marini, A.; García-González, P.; Rubio, A. *Phys. Rev. Lett.* **2006**, *96*, 136404.
- (11) Jiang, H.; Engel, E. *J. Chem. Phys.* **2007**, *127*, 184108.
- (12) Harl, J.; Kresse, G. *Phys. Rev. B* **2008**, *77*, 045136.
- (13) Furche, F. *J. Chem. Phys.* **2008**, *129*, 114105.
- (14) Scuseria, G. E.; Henderson, T. M.; Sorensen, D. C. *J. Chem. Phys.* **2008**, *129*, 231101.
- (15) Toulouse, J.; Gerber, I. C.; Jansen, G.; Savin, A.; Ángyán, J. G. *Phys. Rev. Lett.* **2009**, *102*, 096404.
- (16) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *130*, 081105.

- (17) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 034110.
- (18) Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 154106.
- (19) Ren, X.; Rinke, P.; Scheffler, M. *Phys. Rev. B* **2009**, *80*, 045402.
- (20) Lu, D.; Li, Y.; Rocca, D.; Galli, G. *Phys. Rev. Lett.* **2009**, *102*, 206411.
- (21) Harl, J.; Kresse, G. *Phys. Rev. Lett.* **2009**, *103*, 056401.
- (22) Nguyen, H.-V.; de Gironcoli, S. *Phys. Rev. B* **2009**, *79*, 205114.
- (23) Grüneis, A.; Marsman, M.; Harl, J.; Schimka, L.; Kresse, G. *J. Chem. Phys.* **2009**, *131*, 154115.
- (24) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I. *J. Chem. Theory Comput.* **2010**, *6*, 127–134.
- (25) Nguyen, H.-V.; Galli, G. *J. Chem. Phys.* **2010**, *132*, 044109.
- (26) Hellgren, M.; von Barth, U. *J. Chem. Phys.* **2010**, *132*, 044101.
- (27) Heßelmann, A.; Görling, A. *Mol. Phys.* **2010**, *108*, 359–372.
- (28) Paier, J.; Janesko, B. G.; Henderson, T. M.; Scuseria, G. E.; Grüneis, A.; Kresse, G. *J. Chem. Phys.* **2010**, *132*, 094103.
- (29) Harl, J.; Schimka, L.; Kresse, G. *Phys. Rev. B* **2010**, *81*, 115126.
- (30) Ismail-Beigi, S. *Phys. Rev. B* **2010**, *81*, 195126.
- (31) Zhu, W.; Toulouse, J.; Savin, A.; Ángyán, J. G. *J. Chem. Phys.* **2010**, *132*, 244108.
- (32) Eshuis, H.; Yarkony, J.; Furche, F. *J. Chem. Phys.* **2010**, *132*, 234114.
- (33) Toulouse, J.; Zhu, W.; Ángyán, J. G.; Savin, A. *Phys. Rev. A* **2010**, *82*, 032502.
- (34) Jansen, G.; Liu, R.-F.; Ángyán, J. G. *J. Chem. Phys.* **2010**, *133*, 154106.
- (35) Lu, D.; Nguyen, H.-V.; Galli, G. *J. Chem. Phys.* **2010**, *133*, 154110.
- (36) Heßelmann, A.; Görling, A. *Phys. Rev. Lett.* **2011**, *106*, 093001.
- (37) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I. *J. Chem. Phys.* **2011**, *134*, 114110.
- (38) Ren, X.; Tkatchenko, A.; Rinke, P.; Scheffler, M. *Phys. Rev. Lett.* **2011**, *106*, 153003.
- (39) Lotrich, V.; Bartlett, R. J. *J. Chem. Phys.* **2011**, *134*, 184108.
- (40) Klopper, W.; Teale, A. M.; Coriani, S.; Pedersen, T. B.; Helgaker, T. *Chem. Phys. Lett.* **2011**, *510*, 147–153.
- (41) Heßelmann, A. *J. Chem. Phys.* **2011**, *134*, 204107.
- (42) Kurth, S.; Perdew, J. P. *Phys. Rev. B* **1999**, *59*, 10461–10468.
- (43) Furche, F. *J. Chem. Phys.* **2001**, *114*, 5982–5992.
- (44) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.
- (45) Ángyán, J. G.; Gerber, I. C.; Savin, A.; Toulouse, J. *Phys. Rev. A* **2005**, *72*, 012510.
- (46) Toulouse, J.; Zhu, W.; Savin, A.; Jansen, G.; Ángyán, J. G. *J. Chem. Phys.* **2011**, *135*, 084119.
- (47) Langreth, D. C.; Perdew, J. C. *Solid State Commun.* **1975**, *17*, 1425–1429.
- (48) Langreth, D. C.; Perdew, J. C. *Phys. Rev. B* **1977**, *15*, 2884–2901.
- (49) McLachlan, A. D.; Ball, M. A. *Rev. Mod. Phys.* **1964**, *36*, 844–855.
- (50) Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9677–9686.
- (51) Henderson, T. H.; Scuseria, G. *Mol. Phys.* **2010**, *108*, 2511–2517.
- (52) Oddershede, J. *Adv. Quantum Chem.* **1978**, *11*, 275–352.
- (53) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. MOLPRO, version 2010.2, a package of ab initio programs. See <http://www.molpro.net> (accessed Aug 15, 2011).
- (54) Goll, E.; Werner, H.-J.; Stoll, H.; Leininger, T.; Gori-Giorgi, P.; Savin, A. *Chem. Phys.* **2006**, *329*, 276–282.
- (55) Gerber, I. C.; Ángyán, J. G. *Chem. Phys. Lett.* **2005**, *415*, 100–105.
- (56) Fromager, E.; Toulouse, J.; Jensen, H. J. Aa. *J. Chem. Phys.* **2007**, *126*, 074111.
- (57) Kutzelnigg, W.; Morgan, J., III. *J. Chem. Phys.* **1992**, *96*, 4484–4508. Erratum: **1992**, *97*, 8821.
- (58) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.
- (59) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- (60) Davidson, E. R.; Hagstrom, S. A.; Chakravorty, S. J.; Umar, V. M.; Fischer, C. F. *Phys. Rev. A* **1991**, *44*, 7071–7083.
- (61) Chakravorty, S. J.; Gwaltney, S. R.; Davidson, E. R.; Parpia, F. A.; Fischer, C. F. *Phys. Rev. A* **1993**, *47*, 3649–3670.
- (62) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (63) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (64) Zhao, Q.; Morrison, R. C.; Parr, R. G. *Phys. Rev. A* **1994**, *50*, 2138–2142.
- (65) Boese, A. D. Private communication, 2002. The ZMP potentials have been prepared by the methodology described in refs 68 and 69 using densities calculated with the Brueckner CCD method, obtained with a TZ2P basis (TZ3P for the elements of the third row of the periodic table).
- (66) Gallagher, J.; R.D. Johnson, I. *NIST Chemistry WebBook*, NIST Standard Reference Database Number 69; (Data from ref 70). Linstrom, P. J., Mallard, W. G., Eds.; NIST: Gaithersburg, MD, retrieved in 2010.
- (67) Tang, T.-H.; Toennies, J. P. *J. Chem. Phys.* **2003**, *118*, 4976–49983.
- (68) Tozer, D.; Ingamells, V.; Handy, N. *J. Chem. Phys.* **1996**, *105*, 9200.
- (69) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497.
- (70) Huber, K.; Herzberg, G. *Molecular Spectra and Molecular Structure*; Van Nostrand Reinhold Company: New York, 1979; Vol. IV, Constants of Diatomic Molecules.

Segmented Contracted Douglas–Kroll–Hess Adapted Basis Sets for Lanthanides

Michael Dolg*

Theoretical Chemistry, University of Cologne, Greinstr. 4, 50939 Cologne, Germany

Supporting Information

ABSTRACT: Segmented contracted scalar-relativistic $(23s16p12d6f)/[18s12p9d3f]$ all-electron basis sets for lanthanides La–Lu primarily for use in second-order Douglas–Kroll–Hess density functional calculations are presented. Atomic test calculations at the scalar-relativistic Hartree–Fock level reveal an accurate description of the first to fourth ionization potentials as well as low-energy d–f and d–p excitation energies; i.e., reference data obtained with optimized $(34s28p22d16f)$ even-tempered basis sets are reproduced with mean absolute errors of 0.003 (IP₁), 0.013 (IP₂), 0.030 (IP₃), 0.098 (IP₄), 0.070 (d–f), and 0.018 (d–p) eV. Results of molecular test calculations are presented for the lanthanide trihalides LnX₃ (Ln = La–Lu, X = F, Cl, Br, I) at the PBE0 hybrid density functional theory level. Compared to recently published basis sets of identical size, the sets proposed here show substantially smaller errors in the atomic test calculations as well as lower total energies and produce results of similar accuracy in the molecular calibration study.

1. INTRODUCTION

Several Gaussian basis sets are nowadays available from literature for accurate wave-function-based relativistic all-electron (AE) quantum chemical calculations of lanthanide systems when using the Douglas–Kroll–Hess (DKH) Hamiltonian.^{1–4} Hirao and co-workers optimized $(27s23p15d10f)$ basis sets at the Hartree–Fock (HF) level for Ce to Lu for use with the third-order DKH (DKH3) Hamiltonian, both for a point nucleus as well as for a finite nucleus with a Gaussian charge distribution.^{5,6} The corresponding La basis sets do not contain f functions, i.e., $(27s23p15d)$. Koga and collaborators developed segmented contracted correlating basis sets for Ce to Lu also to be used with the DKH3 Hamiltonian.⁷ The underlying primitive set sizes are $(30s26p23d16f10g10h)$ for Ce to Tb, $(29s25p22d15f9g9h)$ for Dy to Yb, as well as $(28s24p21d14f9g9h)$ for Lu. A corresponding basis set for La is based on a $(30s24p20d15f10g10h)$ primitive set.⁸ Quite recently, Roos et al. published generalized contracted $(25s22p15d11f4g2h)/[12s11p8d7f4g2h]$ atomic natural orbital⁹ (ANO-RCC) basis sets for Ce to Lu for use with the second-order DKH (DKH2) Hamiltonian.¹⁰ The corresponding La basis set is slightly smaller, i.e., $(24s21p15d11f4g)/[11s10p8d5f3g]$. Averaged density matrices for the ground state of the atom and the monocation as well as one excited state of the atom were diagonalized, and the coefficients for all orbitals with occupation numbers larger than 10^{-6} were kept for the contractions. Basis sets of pVXZ quality (X = D, T, Q) as well as minimal basis sets can be derived by omitting a suitable number of least populated ANO contractions in each angular symmetry. For an overview of DKH-adapted basis sets covering also elements other than the lanthanides, the reader is referred to recent review articles by Peterson¹¹ and Nakajima and Hirao.⁴

Recently, Pantazis and Neese proposed loosely segmented contracted $(23s16p12d3f)/[18s12p9d3f]$ (La) and $(23s16p12d6f3g)/[18s12p9d3f3g]$ (Ce–Lu) AE relativistic contracted (SARC) basis sets for the lanthanides to be used in the framework of density

functional theory (DFT) in connection with either the DKH or the zeroth-order regular approximation (ZORA) Hamiltonian.¹² The basis sets were tested for 56 lanthanide trihalides LnX₃ (Ln = Ce–Lu; X = F, Cl, Br, I) at the DKH level using the PBE0 hybrid functional of Adamo and Barone,¹³ which is obtained from the Perdew, Burke, and Ernzerhoff (PBE) gradient corrected density functional^{14,15} by admixing 25% of the exact exchange. The lanthanum trihalides were not considered, but some results for La as well as Lu diatomics LnX (Ln = La, Lu; X = H, O, F) were provided. The authors emphasized that their basis sets, despite their compact size, provide a balanced treatment of different electronic configurations of the lanthanides. Thus, they can be used with confidence for the prediction of energetic properties and provide an unbiased description of processes involving changes in oxidation states and the concomitant changes of the 4f and 5d occupation numbers. Pantazis and Neese also proposed an AE DKH2 or ZORA treatment applying their basis sets as a very efficient alternative to effective core potentials (ECPs) in routine DFT studies of chemically relevant systems.

In the present contribution, we first compare the accuracy of the SARC DKH2 basis sets at the HF level for the first four ionization potentials as well as the low-energy f–d and f–p excitations to results obtained with scalar-relativistic Wood–Boring (WB) adjusted ab initio pseudopotentials (PP) and corresponding standard segmented contracted $(14s13p10d8f6g)/[10s8p5d4f3g]$ valence basis sets^{16,17} as well as with near-HF-limit AE DKH2 and DKH3 results obtained with optimized uncontracted even-tempered $(34s28p22d16f)$ basis sets. Since the results obtained with the SARC basis sets are found to be in an overall lesser agreement with the reference data than the PP results, we present $(23s16p12d6f)/[18s12p9d3f]$ DKH2 basis sets using the same contraction pattern as the SARC sets, but performing

Received: July 22, 2011

Published: August 19, 2011

significantly better and also leading to substantially lower total energies. Second, we present a calibration study for the lanthanide trihalides LnX_3 ($\text{Ln} = \text{La} - \text{Lu}$, $\text{X} = \text{F}, \text{Cl}, \text{Br}, \text{I}$) at the PBE0 density functional theory level and show that the new basis sets lead to equally good results for geometries and atomization energies as the SARC basis sets, despite the absence of very diffuse functions. A brief preliminary report on the performance of the new basis sets in atomic calculations was already given in a discussion of the accuracy and efficiency of density functional calculations applying 4f-in-core PPs on these systems.¹⁸

2. COMPUTATIONAL DETAILS

The HF and DFT calculations reported here were performed with the MOLPRO program system²⁰ using the general-order DKH (DKHn) routines of Hess and co-workers.³ In the atomic HF calculations, symmetry breaking was avoided by calculating a state average including all components of the LS state under consideration by means of the multiconfiguration self-consistent field (MCSCF) code^{21,22} and applying the suitable orbital occupation number restrictions. In order to obtain DKH2 HF reference data, an even-tempered (30s24p18d12f) basis set was optimized by minimizing the sum of the lowest LS state energies of the $4f^{n+1}6s^2$, $4f^n5d^16s^2$, and $4f^n6s^26p^1$ configurations.²³ Since an energy optimization will usually not produce diffuse enough functions, the basis sets were systematically increased by adding n times (1s1p1d1f), keeping the center of gravity as well as the ratio between the exponents fixed. For $n = 4$, the changes in the total energy dropped to below 1 mH (milli-Hartree), and the $f-d$ and $f-p$ excitation energies were typically converged to 0.001 eV. The largest (34s28p22d16f) basis sets thus result by an addition of two tight and two diffuse functions for each angular symmetry to the final (30s24p18d12f) uncontracted even-tempered set. Only this largest set was applied in the evaluation of the DKH2 first to fourth ionization potentials as well as the corresponding DKH3 calculations. The DKH2 ground state energies are found to be slightly lower than the values reported by Pantazis and Neese for their extended universal Gaussian basis sets (34s24p20d14f). Although our total energies are also not fully converged, we think that our ionization potentials and excitation energies are within 0.01 eV of the HF limit.

In order to derive segmented contracted (23s16p12d6f)/[18s12p9d3f] DKH2 basis sets similar to the SARC sets of Pantazis and Neese,¹² a (22s15p11d5f) set was first energy optimized for La–Lu as described above for the even-tempered sets. Thereafter, keeping the (22s15p11d5f) set fixed, one diffuse function for each angular momentum symmetry was optimized in the same manner yielding a (23s16p12d6f) uncontracted set. The derivation of the contraction coefficients from the eigenvectors of various averaged density matrices was explored; however, the adoption of the coefficients for the lowest state of $4f^{n+1}6s^2$ for La–Yb and $4f^n5d^16s^2$ for Lu led to the best overall results. It should be noted that, similar to Pantazis and Neese, the basis sets were derived entirely at the HF level of theory, although they are intended to be later used in DFT calculations. However, in contrast to Pantazis and Neese, La is treated identically to the other lanthanides Ce to Lu, since it is well-known that, although the 4f shell is not occupied in the La ground state, a flexible enough f function set is needed for accurate molecular results.¹⁹

The pseudopotential (PP) calculations reported here for comparison used the WB adjusted energy-consistent PPs of the Stuttgart type^{16,24} and corresponding (14s13p10d8f6g)/[10s8p5d4f3g]

segmented contracted valence basis sets.²⁵ Although the associated (14s13p10d8f6g)/[6s6p5d4f3g] generalized contracted valence basis sets²⁴ are more compact and lead to much better total valence energies, we use the segmented contracted sets here in order to get an unbiased comparison with the AE DKH segmented contracted basis sets.

The molecular calculations on the lanthanide trihalides LnX_3 ($\text{Ln} = \text{La} - \text{Lu}$; $\text{X} = \text{F}, \text{Cl}, \text{Br}, \text{I}$) were performed with the PBE0 hybrid functional of Adamo and Barone,¹³ which is based on the Perdew, Burke, and Ernzerhof (PBE) gradient corrected functional.¹⁴ This functional was also used in the study of Pantazis and Neese and turned out to give the overall best results in our recent PP study of the lanthanide trihalides.^{18,26} Since the lanthanum trihalides have not been included in the study by Pantazis and Neese, the corresponding calculations were carried out also using their La basis set. However, since their basis set does not provide enough f functions, the corresponding results are not included when calculating the mean absolute deviations between theoretical and recommended or experimental reference data. The basis sets applied for the halides, i.e., relativistically recontracted polarized triple- ζ basis sets, were taken from the ORCA basis set library.²⁷ It should be noted that these basis sets only keep the innermost contraction for s and p symmetry and thus are in fact larger than their nonrelativistic polarized triple- ζ counterparts. The DKH2 sets denoted as TZVP in the ORCA basis set library, i.e., F (11s6p1d)/[6s3p1d], Cl (14s9p1d)/[8s4p1d], Br (17s13p7d)/[10s8p3d], and I (19s15p10d)/[12s10p6d], were applied in DFT calculations using the ORCA code,²⁷ whereas extended DKH2 sets denoted as TZVPP, i.e., F (11s6p2d1f)/[6s3p2d1f], Cl (14s9p2d1f)/[8s4p2d1f], Br (17s13p8d1f)/[10s8p4d1f], and I (19s15p11d2f)/[12s10p7d2f], were applied in calculations using MOLPRO.^{20,28} For the lanthanides, the (23s16p12d6f)/[18s12p9d3f] sets presented here as well as the SARC sets of Pantazis and Neese were used. A g set was not added, since it was found to lead only to slight changes at the DFT level in test calculations; i.e., the bond lengths and the atomization energies of the lanthanide triiodides LnI_3 ($\text{Ln} = \text{La} - \text{Lu}$) changed on average by 0.003 Å and 0.04 eV, respectively. These values are far below the experimental errors bars and most likely smaller than errors introduced by the usage of DFT.

The results reported here refer to the unrestricted Kohn–Sham formalism and a single Slater determinant as a model wave function. The determinant leading to the lowest total energy was searched for by applying successively all possible rotations between the differently occupied 4f orbitals. The DFT calculations were performed in C_1 symmetry; i.e., all orbital rotations and mixings were allowed by symmetry. However, except as otherwise noted, C_{3v} symmetry was imposed on the molecular structure during the optimization.²³ The target accuracy for integration within the MOLPRO DFT module^{20,28} was changed to 10^{-11} or better, compared to a default value of 10^{-6} . The calculations with the SARC basis sets were repeated here with the ORCA DFT module,²⁷ since from the publication of Pantazis and Neese¹² it is not entirely clear how the open f shells were treated and to which atomic and molecular states or averages the atomization energies refer.^{29,30}

The destabilization of chemical bonds by spin–orbit (SO) splitting is a well-known effect for heavy element molecules,^{31,32} since typically the atomic total energy SO lowerings are larger than the molecular ones. For halogen compounds, the SO induced reductions of the atomization energies might roughly amount to the values one may estimate from the experimental

Table 1. Ground State DKH2 Hartree–Fock Energies (Hartree) Obtained with the Proposed Basis Sets of This Work in Comparison to the SARC Basis Sets of Pantazis and Neese¹² and Near HF-Limit Values Obtained with Optimized Uncontracted Even-Tempered (34s28p22d16f) Sets

Ln	Ln	this work		SARC		near HF
		<i>E</i>	error	<i>E</i>	error	<i>E</i>
La	5d ¹ 6s ² 2D	−8486.3096	0.1109	−8485.4449	0.9756	−8486.4205
Ce	4f ¹ 5d ¹ 6s ² 3F	−8853.0578	0.2367	−8852.0936	1.2009	−8853.2945
Pr	4f ³ 6s ² 4I	−9229.4752	0.2729	−9228.3765	1.3716	−9229.7481
Nd	4f ⁴ 6s ² 5I	−9615.7139	0.2979	−9614.4993	1.5125	−9616.0118
Pm	4f ⁵ 6s ² 6H	−10011.8533	0.3276	−10010.5149	1.6660	−10012.1809
Sm	4f ⁶ 6s ² 7F	−10418.0751	0.3617	−10416.6038	1.8330	−10418.4368
Eu	4f ⁷ 6s ² 8S	−10834.5405	0.4029	−10832.9265	2.0169	−10834.9434
Gd	4f ⁷ 5d ¹ 6s ² 9D	−11261.2456	0.4485	−11259.5479	2.1462	−11261.6941
Tb	4f ⁹ 6s ² 6H	−11698.1351	0.4886	−11696.1825	2.4412	−11698.6237
Dy	4f ¹⁰ 6s ² 5I	−12145.8345	0.5376	−12143.6953	2.6768	−12146.3721
Ho	4f ¹¹ 6s ² 4I	−12604.2724	0.5914	−12601.9349	2.9289	−12604.8638
Er	4f ¹² 6s ² 3H	−13073.6050	0.6408	−13071.0476	3.1982	−13074.2458
Tm	4f ¹³ 6s ² 2F	−13554.0235	0.7054	−13551.2397	3.4892	−13554.7289
Yb	4f ¹⁴ 6s ² 1S	−14045.7231	0.7696	−14042.6902	3.8025	−14046.4927
Lu	4f ¹⁴ 5d ¹ 6s ² 2D	−14548.7669	0.8521	−14545.6043	4.0147	−14549.6190
m.a.e.			0.4696		2.3516	

atomic ${}^2P_{3/2}$ – ${}^2P_{1/2}$ SO splittings,³³ i.e., LnF₃, −0.050; LnCl₃, −0.109; LnBr₃, −0.456; and LnI₃, −0.942 eV. The situation however is more complex due to the atomic-like character the open Ln 4f shell retains in the LnX₃ molecules. In those few cases where the 4f occupation number is the same in the free Ln atom and the LnX₃ molecule, the contributions from the 4f shell may approximately cancel, and a further SO induced reduction of the atomization energies only results from the atomic SO splitting of the 5d shells, e.g., for LaX₃, CeX₃, GdX₃, and LuX₃ (X = F, Cl, Br, I). However, the majority of cases involve different 4f occupations in the Ln atom and the LnX₃ molecule. The 4f SO contributions may not cancel and cannot be simply estimated on the basis of atomic data. Therefore, atomic and molecular SO contributions to the total energy were calculated using the MOLPRO code^{20,34} at the DKH2 RHF level using the Breit–Pauli Hamiltonian in quasi-degenerate first-order perturbation theory. To the best of our knowledge, such corrections have not been included in previous studies of the lanthanide trihalides,^{12,35–41} except for a relativistic AE DFT study by Adamo and Maldivi.⁴² These corrections of the atomization energies were only derived for the basis sets proposed here; however, they were also applied to the scalar-relativistic DKH2 results obtained with the Ln basis sets of Pantazis and Neese¹² for comparison to experimental data.

3. RESULTS AND DISCUSSION

The following section summarizes results for atomic test calculations at the HF level and molecular test calculations on the lanthanide trihalides LnX₃ (Ln = La–Lu; X = F, Cl, Br, I) at the DFT level. For a few cases, results of HF and coupled cluster calculations with single, double, and perturbative triple excitations (CCSD(T)) performed with the MOLPRO code^{20,43,44} are also reported for comparison.

3.1. Atomic Calculations. One criterion on which to judge the quality of a basis set is the total energy. The ground state HF energies derived with the proposed basis sets are between 0.87

(La) and 3.16 (Lu) Hartree lower than the energies obtained with the SARC basis sets,¹² cf. Table 1. The total ground state energies of the new sets are still at least between 0.11 (La) and 0.85 (Lu) Hartree above the HF limit (mean absolute error (m.a.e.) 0.47 hartree), which is most likely close to the values obtained with our optimized uncontracted even-tempered (34s28p22d16f) sets. The corresponding errors of the sets of Pantazis and Neese are between 0.98 (La) and 4.01 (Lu) Hartree (m.a.e. 2.35 hartree). Our estimates of the DKH2 HF limit are for all 15 lanthanide atoms lower than the values obtained by Pantazis and Neese with (34s24p20d14f) sets based on the universal Gaussian basis sets (UGBS) of de Castro et al.^{45,46} Our HF results for the SARC basis sets in some cases deviate from the spin-averaged restricted open-shell HF (ROHF) values published by Pantazis and Neese.¹² We can only speculate about the reasons and want to point out that in our calculations symmetry breaking is avoided; i.e., all components of the LS state under consideration are calculated and are degenerate. Moreover, in corresponding nonrelativistic calculations, we approach finite difference HF results obtained with the atomic code MCHF77⁴⁷ for the LS state.

The open 4f shell causes various difficulties in calculations on lanthanide systems, e.g., large differential relativistic effects usually destabilizing states with high 4f occupation and large counteracting correlation effects stabilizing states with high 4f occupation have to be dealt with.^{48,49} Such problems become relevant, e.g., for the calculation of the LnX₃ atomization energies, since in these molecules according to the ionic charge distribution Ln³⁺(X[−])₃ the Ln³⁺ centers have a 4f^{*n*} configuration (*n* = 0–14 for La–Lu), whereas in the ground states of the neutral Pr–Eu and Tb–Yb atoms one has a 4f^{*n*+1}6s² configuration.⁵⁰ Thus, only for La, Ce, Gd, and Lu, where the neutral atoms have a 4f^{*n*}5d¹6s² configuration, may one expect a rough cancellation of differential relativistic and correlation effects in the calculation of the atomization energies.

Aside from questions of the accuracy of the relativistic Hamiltonian itself and the applied density functional for the evaluation of 4f^{*n*+1}6s²–4f^{*n*}5d¹6s² energy differences, it is also

Table 2. First to Fourth AE DKH2 Ionization Potentials, $4f^n 5d^1 6s^2 \rightarrow 4f^{n+1} 6s^2$ and $4f^n 5d^1 6s^2 \rightarrow 4f^n 6s^2 6p^1$ Excitation Energies Obtained at the HF Level with the (34s28p22d16f) Basis Sets (DKH2 Limit), and the (23s16p12d6f)/[18s12p9d3f] Segmented Contracted Sets of This Work As Well As the SARC Basis Sets of Pantazis and Neese^{a,b,12}

Ln	IP ₁			IP ₂			IP ₃		
	DKH2 limit	this work	SARC	DKH2 limit	this work	SARC	DKH2 limit	this work	SARC
La	4.37	4.37	4.40	10.36	10.36	10.35	18.10	18.10	18.07
Ce	4.56	4.55	4.57	11.14	11.30	11.47	18.24	18.11	17.97
Pr	4.43	4.43	4.44	9.89	9.89	9.91	19.68	19.60	19.39
Nd	4.47	4.47	4.48	10.04	10.04	10.06	19.81	19.75	19.67
Pm	4.52	4.52	4.52	10.19	10.19	10.21	19.83	19.80	19.80
Sm	4.56	4.56	4.56	10.34	10.34	10.36	21.05	21.03	21.07
Eu	4.59	4.59	4.60	10.48	10.47	10.50	22.73	22.72	22.78
Gd	4.79	4.79	4.80	11.36	11.37	11.38	19.51	19.51	19.51
Tb	4.76	4.76	4.76	10.71	10.71	10.73	17.62	17.65	17.84
Dy	4.83	4.83	4.84	10.82	10.82	10.84	18.85	18.87	19.07
Ho	4.91	4.91	4.92	10.93	10.93	10.95	18.50	18.48	18.74
Er	4.99	4.98	5.00	11.04	11.04	11.07	18.10	18.08	18.35
Tm	5.07	5.06	5.07	11.16	11.15	11.18	19.21	19.20	19.47
Yb	5.14	5.13	5.15	11.26	11.26	11.28	20.92	20.90	21.16
Lu	4.35	4.35	4.34	12.59	12.59	12.61	19.94	19.94	19.97
MAD	0.000	0.003	0.009	0.000	0.013	0.041	0.000	0.030	0.153
	0.003	0.006	0.005	0.006	0.016	0.035	0.009	0.027	0.155

Ln	IP ₄			d–f			d–p		
	DKH2 limit	this work	SARC	DKH2 limit	this work	SARC	DKH2 limit	this work	SARC
La	49.89	49.90	49.89	2.74	2.96	a)	1.86	1.84	1.81
Ce	34.90	34.67	34.30	1.34	1.48	1.63	2.12	2.13	2.13
Pr	37.23	37.11	36.76	0.11	0.21	0.48	2.10	2.12	2.15
Nd	38.93	38.88	38.82	0.26	0.33	0.48	2.21	2.24	2.27
Pm	39.17	39.18	39.40	0.52	0.57	0.64	2.27	2.30	2.31
Sm	39.31	39.37	39.79	−0.47	−0.43	−0.41	2.26	2.30	2.29
Eu	40.81	40.91	41.42	−2.09	−2.07	−2.09	2.09	2.14	2.12
Gd	42.83	42.94	42.93	4.68	4.67	5.14	1.72	1.74	1.72
Tb	35.25	35.42	36.45	3.24	3.23	3.06	1.87	1.89	1.88
Dy	37.46	37.63	38.62	2.11	2.09	1.93	1.79	1.81	1.79
Ho	38.95	39.06	40.10	2.11	2.08	1.89	1.32	1.33	1.31
Er	38.73	38.83	39.92	2.42	2.36	2.18	1.05	1.05	1.03
Tm	38.45	38.57	39.66	1.61	1.57	1.38	1.14	1.14	1.11
Yb	39.84	39.95	41.01	0.10	0.05	−0.13	1.12	1.12	1.10
Lu	41.85	41.85	42.18				0.30	0.30	0.30
MAD	0.000	0.098	0.667	0.000	0.070	0.215	0.000	0.018	0.024
	0.008	0.103	0.673	0.008	0.064	0.218	0.005	0.020	0.023

^a The results are for the lowest LS states (cf. Supporting Information). The mean absolute deviations (MAD) listed in the first line at the bottom refer to the AE DKH2 near HF results from this table, those in the second line to corresponding DKH3 data from Table 3. ^b Value 10.66 eV not included in MAD

important to assess the quality of the applied basis set. Tables 2 and 3 present DKH2 and DKH3 AE HF results, respectively, obtained with the large optimized uncontracted even-tempered (34s28p22d16f) basis sets for ionization potentials and excitation energies, which are assumed to be quite close to the basis set limit. Reduced basis sets containing one or two primitives for each angular quantum number less virtually give the same values. The differences between the DKH2 and DKH3 results are quite small, and it thus may be hoped that the DKH2 Hamiltonian provides already a quite good scalar-relativistic description for lanthanide systems.

Table 2 also lists the DKH2 HF results obtained with the SARC basis sets of Pantazis and Neese.¹² The errors are quite small for the first and second ionization potentials, but they amount to several tenths of an electronvolt for the third and fourth ionization potentials (mean absolute deviations (MAD) of 0.01, 0.04, 0.15, and 0.67 eV for IP₁, IP₂, IP₃, and IP₄, respectively), especially in the cases where the 4f occupation is changing. These results can be compared to those obtained with the WB-adjusted small-core PPs^{16,24} and their corresponding standard segmented contracted (14s13p10d8f)/[10s8p5d4f] basis sets²⁵ listed in Table 3 (MAD = 0.04, 0.13, 0.11, and 0.32 eV).

Table 3. First to Fourth AE DKH3 Ionization Potentials, $4f^n 5d^1 6s^2 \rightarrow 4f^{n+1} 6s^2$ and $4f^n 5d^1 6s^2 \rightarrow 4f^n 6s^2 6p^1$ Excitation Energies Obtained at the HF Level Obtained with the (34s28p22d16f) Basis Sets (DKH3 Limit), and Corresponding Small-Core PP¹⁶ Results Using the (14s13p10d8f)/[10s8p5d4f] Segmented²⁴ As Well As the (14s13p10d8f)/[6s6p5d4f] Generalized²⁵ Valence Basis Sets^a

Ln	IP ₁			IP ₂			IP ₃		
	DKH3	PP	PP	DKH3	PP	PP	DKH3	PP	PP
	limit	seg.	gen.	limit	seg.	gen.	limit	seg.	gen.
La	4.38	4.40	4.39	10.35	10.35	10.36	18.10	18.06	18.06
Ce	4.56	4.62	4.62	11.15	11.18	11.20	18.23	18.29	18.29
Pr	4.43	4.47	4.47	9.89	9.97	9.97	19.67	19.80	19.80
Nd	4.47	4.51	4.52	10.04	10.13	10.13	19.80	19.94	19.94
Pm	4.52	4.55	4.55	10.19	10.26	10.26	19.82	19.97	19.96
Sm	4.56	4.58	4.58	10.34	10.38	10.38	21.04	21.19	21.19
Eu	4.60	4.61	4.61	10.48	10.50	10.50	22.72	22.86	22.86
Gd	4.80	4.86	4.86	11.37	11.48	11.49	19.50	19.74	19.74
Tb	4.76	4.71	4.72	10.72	10.62	10.61	17.61	17.83	17.82
Dy	4.84	4.81	4.81	10.83	10.78	10.77	18.84	19.05	19.03
Ho	4.91	4.90	4.91	10.94	10.92	10.92	18.49	18.60	18.60
Er	4.99	4.97	4.97	11.05	11.00	11.00	18.09	18.21	18.21
Tm	5.07	5.03	5.03	11.16	11.09	11.08	19.21	19.30	19.30
Yb	5.14	5.06	5.07	11.27	11.11	11.11	20.91	20.92	20.93
Lu	4.34	4.38	4.38	12.60	12.67	12.67	19.95	19.97	19.94
MAD	0.003	0.037	0.035	0.073	0.125	0.128	0.009	0.113	0.111
	0.000	0.036	0.035	0.000	0.065	0.069	0.000	0.121	0.119

Ln	IP ₄			d–f			d–p		
	DKH3	PP	PP	DKH3	PP	PP	DKH3	PP	PP
	limit	seg.	gen.	limit	seg.	gen.	limit	seg.	gen.
La	49.89	49.94	49.94	2.74	2.67	2.68	1.83	1.84	1.85
Ce	34.89	35.09	35.09	1.34	1.34	1.35	2.12	2.18	2.19
Pr	37.23	37.50	37.50	0.11	0.09	0.10	2.10	2.20	2.21
Nd	38.92	39.23	39.23	0.27	0.24	0.25	2.21	2.31	2.32
Pm	39.16	39.50	39.49	0.53	0.49	0.50	2.27	2.36	2.37
Sm	39.30	39.65	39.64	−0.46	−0.53	−0.52	2.25	2.31	2.32
Eu	40.80	41.14	41.12	−2.08	−2.17	−2.16	2.08	2.13	2.14
Gd	42.82	43.48	43.46	4.69	4.65	4.68	1.72	1.86	1.87
Tb	35.24	35.69	35.69	3.25	3.10	3.13	1.87	1.87	1.88
Dy	37.46	37.92	37.90	2.12	1.97	2.01	1.78	1.80	1.82
Ho	38.94	39.35	39.33	2.12	2.01	2.04	1.32	1.34	1.35
Er	38.72	39.15	39.12	2.43	2.28	2.30	1.05	1.04	1.05
Tm	38.45	38.80	38.77	1.62	1.46	1.49	1.13	1.10	1.12
Yb	39.83	40.04	40.03	0.11	−0.07	−0.05	1.12	1.00	1.03
Lu	41.83	41.91	41.91				0.30	0.30	0.35
MAD	0.008	0.319	0.308	0.008	0.082	0.064	0.005	0.053	0.059
	0.000	0.327	0.316	0.000	0.090	0.072	0.000	0.054	0.061

^a A diffuse p function was added to the PP basis sets for the evaluation of d–p excitation energies. The results are for the lowest LS states (cf. supplementary material). The mean absolute deviations (MAD) listed in the first line at the bottom refer to the AE DKH2 near HF results from Table 2, those in the second line to corresponding DKH3 data from this table.

Essentially identical results are obtained for the small-core PPs when the more compact generalized contracted (14s13p10d8f)/[6s6p5d4f] basis sets are used (MAD = 0.04, 0.13, 0.11, and 0.32 eV). We note here in passing that the HF PP results for IP₁ of La, Ce, Gd, and Lu in Table 2 are by 0.2–0.4 eV lower than the CAS PP results reported previously,^{24,25} where the 6s shell was included together with the 5d and 4f shells in the active space.

Although the SARC basis sets perform slightly better for the first and second ionization potentials, and slightly worse for the third one, they exhibit substantially larger errors for the fourth ionization potentials, which involve a change of the 4f occupation number for all elements except for La. These errors are most likely not due to the contraction, since the uncontracted SARC sets exhibit different but not really better results. We note that the SARC sets are characterized by the presence of unusually diffuse outermost exponents, and that their exponents have been simply adopted from universal Gaussian basis sets optimized by Jorge et al. in Dirac–Hartree–Fock (DHF) calculations for the Dirac–Coulomb Hamiltonian (DC)^{45,46} instead of optimizing them for the DKH2 or ZORA target Hamiltonian. It is thus not surprising that the tighter exponents describing inner regions of the atom, where (upper component) DHF/DC and DKH2 or ZORA spinors certainly differ in shape, are not optimal. The seemingly favorable DKH2 SARC ionization energies presented by Pantazis and Neese¹² were obtained at the Becke exchange and Lee, Young, Parr correlation hybrid functional (B3LYP) level and compared to experimental values. In view of the quite large error bars for the higher ionization potentials, the omission of SO contributions, and the probably not too accurate description of 4f shell correlation effects by B3LYP and a single reference DFT ansatz, it is thus quite difficult to come to a conclusion on possible basis set errors on the basis of their results.

Similarly, for d–f and d–p excitation energies, the SARC HF DKH2 results (MAD = 0.22 and 0.02 eV) are not clearly better than the PP HF results (MAD = 0.08 and 0.05 eV). As for the ionization potentials, the performance of the segmented²⁵ and generalized²⁴ contracted basis sets is rather similar. It has to be noted, however, that a diffuse p function had to be added to the PP basis sets in order to get a reasonable description of the 6p shell, i.e., reliable d–p excitation energies, whereas the addition of this function changed the calculated ionization potentials and d–f excitation energies only negligibly.

In contrast to this, the optimized DKH2 basis sets proposed here exhibit clearly the best overall performance (MAD = 0.00, 0.01, 0.03, 0.10, 0.07, and 0.02 eV for IP₁, IP₂, IP₃, IP₄, d–f, and d–p, respectively). Although the diffuse outermost exponents in each angular symmetry are somewhat larger than in the SARC sets, they are able to describe quite well the configurations with relatively diffuse valence orbitals, e.g., $4f^{n+1} 6s^2$, $4f^n 5d^1 6s^2$, and $4f^n 6s^2 6p^1$. Since in most molecules the lanthanides are charged Ln^{x+} ($x = 3$ and also 2 or 4) ions, we expect the present basis sets to be sufficiently accurate. An addition of diffuse functions might be considered for calculations of weakly interacting systems, e.g., dimers with van der Waals bonding contributions such as Yb₂, but in these cases, the application of DFT, for which the basis sets are designed, is questionable.

A brief comment on the performance of the small-core PPs¹⁶ is in order. The reasons for the errors are most likely 2-fold. On one hand, only Ln and Ln⁺ configurations, excluding those with an occupied 6p shell, have been considered in the adjustment; i.e., only IP₁ and the d–f excitation energy have been included in the reference data set. On the other hand, the reference energies were

Table 4. Mean Absolute Deviations (cm^{-1}) in Relative Energies of J Levels with Respect to the Parent $\text{Ln}^{3+} 4f^n$ LS Ground State^a

Ln		AE DKH2/BP		PP	
		this work	SARC	seg.	gen.
Ce	$4f^1 2F$	94.5	231.1	218.8	210.6
Pr	$4f^2 3H$	106.4	341.3	251.3	244.6
Nd	$4f^3 4I$	119.0	383.5	254.4	248.7
Pm	$4f^4 5I$	119.7	381.4	223.9	220.6
Sm	$4f^6 6H$	104.7	327.7	157.2	155.2
Eu	$4f^7 7F$	78.9	242.0	90.2	89.1
Tb	$4f^8 7F$	92.3	299.0	296.8	297.2
Dy	$4f^9 6H$	140.1	447.9	418.5	420.7
Ho	$4f^{10} 5I$	160.9	569.4	485.3	490.5
Er	$4f^{11} 4I$	172.4	620.2	266.5	272.8
Tm	$4f^{12} 3H$	162.8	560.5	15.9	29.6
Yb	$4f^{13} 2F$	119.7	393.3	250.2	250.6
avg. MAD		122.6	399.8	244.1	244.2

^aThe reference values were obtained by CASSCF DKH2 calculations using (34s28p22d16f) basis sets and applying the BP Hamiltonian in first-order quasidegenerate perturbation theory. The corresponding results obtained with the segmented contracted (23s16p12d6f)/[18s12p9d3f] basis sets of this work and the corresponding SARC basis sets¹² are compared to results obtained with small-core PPs¹⁶ and (14s13p10d8f)/[10s8p5d4f] segmented²⁵ as well as (14s13p10d8f)/[6s6p5d4f] generalized²⁴ contracted basis sets.

derived with the scalar-relativistic WB HF formalism, which leads to slightly different results from those obtained with the DKH2 Hamiltonian. We note that the PPs were generated more than two decades ago and by now do not correspond to the state of the art. Modern energy-consistent PPs are adjusted to reference data going beyond the WB, DKH2, or ZORA Hamiltonians in accuracy,^{51–56} i.e., they are based on multireference Dirac–Hartree–Fock data obtained with the Dirac–Coulomb–Breit finite nucleus Hamiltonian and possibly also include higher-order corrections from quantum electrodynamics. Such potentials, e.g., the one recently published for the U atom,^{57,58} are certainly in terms of accuracy more competitive with AE calculations than the WB-adjusted PPs.

Spin–orbit (SO) effects cannot be neglected for lanthanide systems (*vide infra*), especially for energy differences between states with different 4f and/or 5d occupations.^{30,59} Table 4 lists mean absolute deviations (MAD) in the relative energy of the J-levels of the Ln^{3+} ions with respect to the parent LS ground state configuration. The Breit–Pauli (BP) Hamiltonian was used in first-order quasidegenerate perturbation theory.²⁰ The average MAD for the 12 elements listed in Table 4 amounts to 400 cm^{-1} for the SARC basis sets, compared to 244 cm^{-1} for both the segmented and generalized contracted PP results. Again, the present AE basis sets perform best, i.e., the average MAD amounts to only 123 cm^{-1} . We assume that by optimizing the even-tempered basis sets underlying the segmented contracted sets proposed here, a better description of the compact 4f shell was obtained than is the case for the SARC basis sets. Detailed results for the individual J levels are available in the Supporting Information.

The number of s, p, d, and f primitives in the extended PP basis sets is 162, compared to 173 in the AE SARC basis sets and the

Table 5. Mean Absolute Errors in PBE0 Calculated Ln–X bond Lengths (Å) of LnX_3 with Respect to Recommended Values by Kovacs and Konings (KK)⁶³ and Hargittai (H)⁶⁴

method	LnF_3		LnCl_3		LnBr_3		LnI_3	
	KK	H	KK	H	KK	H	KK	H
AE DKH ^d	0.016	0.017	0.018	0.021	0.017	0.022	0.023	0.024
PP+CPP ^b	0.006	0.028	0.017	0.023	0.013	0.024	0.027	0.016
AE DKH ^c	0.015	0.020	0.019	0.022	0.022	0.017	0.044	0.014
AE DKH ^d	0.018	0.017	0.021	0.020	0.021	0.018	0.021	0.010
AE DKH ^c	0.015	0.020	0.019	0.022	0.019	0.020	0.046	0.012

^aRef 12, including one-center approximation. ^b4f-in-core PP+CPP results.¹⁸ ^cThis work using the proposed Ln basis sets and X TZVPP basis sets, MOLPRO.²⁰ ^dThis work using the SARC basis sets of ref 12 and X TZVP basis sets, ORCA.²⁷ ^eThis work using the proposed Ln basis sets and X TZVP basis sets, ORCA.²⁷

sets proposed here. The number of s, p, d, and f contractions is 90 and 80 for the segmented and generalized ANO contracted PP basis sets, respectively, compared to 120 for the AE basis sets. The PP basis sets appear to be relatively large; however, they have the advantage that they can also be used in wave-function-based correlation calculations without suffering from too large basis set superposition errors, which is not the case for the small AE basis sets discussed here. In addition, the PP ANO contracted PP sets of VQZ quality can be easily reduced and yield sets with fewer s, p, d, and f contractions, i.e., VTZ (64) and VDZ (48). Note that these numbers always include the diffuse p function, which might also be omitted for most purposes. The construction of more compact segmented contracted PP basis sets for usage in DFT calculations is currently underway in the group of Weigend and will also cover the lanthanides.^{60–62} The segmented contracted AE DKH sets presented here and by Pantazis and Neese¹² have nearly as many s, p, d, and f contractions as the most extended ANO contracted DKH2 sets published by Roos and co-workers,¹⁰ which consist of 134 s, p, d, and f contractions. Again, the advantage of the ANO sets is that smaller basis sets can easily be constructed by omitting the least populated ANO contractions.

3.2. Molecular Results. Molecular test calculations were performed for the lanthanide trihalides LnX_3 (Ln = La–Lu; X = F, Cl, Br, I). An overview over the various previous theoretical studies of these systems^{12,35–42} up to 2004 was given by Kovacs and Konings.⁶³ In the present study, we compare mainly to the recent DKH AE PBE0 results of Pantazis and Neese.¹² A report on the performance of 4f-in-core PPs combined with effective core-polarization potentials (CPPs) in various DFT and CCSD(T) calculations will be given elsewhere.²⁶

3.2.1. LnX_3 Bond Lengths and Angles. A summary of the overall performance of various approaches for LnX_3 bond lengths is listed in Table 5. Detailed geometry parameters of all calculations of this work are available from the Supporting Information. We make comparison to two reference data sets, i.e. the one published by Hargittai⁶⁴ and the one of Kovacs and Konings.⁶³ The latter set usually has shorter bond distances than the former, with the calculated results falling usually in between. In principle, a careful analysis of the recommended values and the calculated results should be carried out for each molecule, especially for those cases where irregularities or surprising results are present, e.g., for YbBr_3 and LuBr_3 . Hargittai lists quite different values of 2.571 and 2.506 Å, respectively, whereas Kovacs and Konings

Table 6. Ln–X Bond Lengths (Å) from Spin-Free AE DKH2 PBE0 Calculations Using the Ln (23s16p12d6f)/[18s12p9d3f] Basis Sets Proposed in This Work As Well As the Corresponding SARC Basis Sets¹² Together with F (11s6p1d)/[6s3p1d], Cl (14s9p1d)/[8s4p1d], Br (17s13p7d)/[10s8p3d], and I (19s15p10d)/[12s10p6d] Basis Sets (Denoted by TZVP in the ORCA Basis Set Library) for the Halides

Ln	LnF ₃		LnCl ₃		LnBr ₃		LnI ₃	
	this work	SARC	this work	SARC	this work	SARC	this work	SARC
La	2.107	2.114	2.576	2.581	2.729	2.735	2.952	2.957
Ce	2.085	2.090	2.554	2.555	2.704	2.704	2.926	2.926
Pr	2.070	2.074	2.535	2.535	2.683	2.685	2.906	2.905
Nd	2.068	2.071	2.519	2.520	2.669	2.669	2.890	2.890
Pm	2.052	2.055	2.506	2.506	2.655	2.654	2.875	2.876
Sm	2.037	2.040	2.493	2.494	2.643	2.645	2.872	2.873
Eu	2.028	2.030	2.487	2.489	2.649	2.651	2.921	2.915
Gd	2.029	2.029	2.470	2.472	2.617	2.619	2.835	2.838
Tb	2.012	2.013	2.456	2.458	2.602	2.604	2.821	2.824
Dy	2.003	2.004	2.442	2.444	2.589	2.592	2.808	2.812
Ho	1.999	2.000	2.431	2.434	2.578	2.580	2.796	2.800
Er	1.986	1.988	2.420	2.424	2.566	2.569	2.785	2.788
Tm	1.980	1.981	2.408	2.413	2.556	2.559	2.775	2.780
Yb	1.966	1.967	2.399	2.403	2.547	2.552	2.770	2.780
Lu	1.963	1.965	2.389	2.391	2.535	2.538	2.754	2.756
MAD	0.002		0.002		0.002		0.003	

give 2.529 and 2.516 Å, respectively. Such a discussion of individual molecules is beyond the scope of the present work.

One notes that large-core pseudopotentials (PPs), when augmented by core-polarization potentials (CPPs) on the lanthanides and halides, lead to results of similar overall quality as AE DKH2 calculations applying the same density functional.¹⁸ As discussed in more detail elsewhere, the lanthanide PPs include the open 4f shell in the core and require significantly smaller computational resources than the AE DKH2 calculations reported here.¹⁹ They also avoid several difficulties which arise when applying density functional theory to open f shell systems, which actually would require a multireference treatment. It is also obvious that the differences between the results obtained with the basis sets proposed here and those of ref 12 are quite small, and even larger deviations arise from the usage of different program systems. This point will be discussed below in more detail.

Table 6 provides a comparison of AE DKH PBE0 bond lengths obtained using the MOLPRO code²⁰ and the basis sets proposed here as well as the SARC basis sets.¹² It is seen that the agreement is excellent with mean average deviations of 0.003 Å or less for all four halides. Quite good agreement is also obtained when comparing to the results published by Pantazis and Neese¹² for the trifluorides and trichlorides, whereas the bond lengths of the tribromides and triiodides obtained here are slightly longer. The main reason for these deviations appears to be the usage of a one-center approximation in the ORCA code²⁷ applied by Pantazis and Neese for the evaluation of the DKH2 contributions in geometry optimizations. This computational detail, which is not mentioned in the original work, was later pointed out by the authors.⁶⁵ Corresponding results of calculations using the ORCA software,²⁷ i.e., applying the one-center approximation,

are summarized in the Supporting Information. Here again, excellent agreement is obtained between the results for both basis sets; i.e., the MAD in bond lengths is 0.003, 0.002, 0.002, 0.011, and 0.004 Å for LnF₃, LnCl₃, LnBr₃, LnI₃, and all LnX₃, respectively. When avoiding the one-center approximation by using ORCA for single-point calculations and an external geometry optimizer, the agreement is even better; i.e., the MAD is 0.002, 0.002, 0.002, 0.003, and 0.003 Å for LnF₃, LnCl₃, LnBr₃, LnI₃, and all LnX₃, respectively.

It has to be noted that both the bond angles and bond lengths published by Pantazis and Neese¹² are in fact averaged values, as it was later revealed by the authors.⁶⁵ Giving just average values is valid for bond lengths, which show variations of less than 0.01 Å; however, for some molecules, especially the europium trihalides, the mere stating of an average bond angle leads to the wrong impression that the molecules in the calculations have either C_{3v} or D_{3h} symmetry. In fact, quite noticeable distortions from these ideal symmetries may occur when imposing no symmetry restrictions in the calculation, and the X–Ln–X bond angles for some systems show mean absolute deviations of up to 8°. For example, for EuI₃, bond angles of 116.2, 116.2, and 127.6° and 116.4, 116.4, and 127.1° were obtained when using the SARC basis sets and the basis sets proposed here, respectively, cf. Supporting Information. The corresponding bond distances are 2.909, 2.914, and 2.914 Å and 2.885, 2.891, and 2.891 Å, respectively. At the multiconfiguration Hartree–Fock level, such distortions do not occur.

Table 7 summarizes the AE DKH PBE0 results obtained with the proposed basis sets and extended sets for the halogen atoms. The comparison is made to an average of the values recommended by Hargittai⁶⁴ and Kovacs and Konings,⁶³ i.e. $R_{\text{avg}} = (R_{\text{H}} + R_{\text{KK}})/2$. The values in parentheses denote the deviations of these recommended values from the average value in the last two printed digits, i.e., $n = 500|R_{\text{H}} - R_{\text{KK}}|$. It should be noted that $R_{\text{H}} > R_{\text{KK}}$, except for LuBr₃. The mean absolute deviations of the AE DKH PBE0 results from the averaged reference values is 0.006 Å for LnF₃, LnCl₃, and LnBr₃ but 0.022 Å for LnI₃. The major part of the deviations arises for the lighter lanthanide iodides; i.e., mean absolute deviations of 0.040 and 0.007 Å are observed for LaI₃ to EuI₃ and GdI₃ to LuI₃, respectively. In all cases, except for LuBr₃, the differences between the values of the two recommended sets are larger than between the calculated results and the values of any of the two sets. This indicates a need for relativistic higher-level correlated ab initio AE calculations to generate suitable reference data for calibration purposes.

The limited accuracy of a density functional description for lanthanide open f shell systems becomes especially apparent for the europium trihalides. The f population increases considerably above the value of 6 electrons at the density functional level, i.e., 6.24, 6.23, 6.30, and 6.56 for EuF₃, EuCl₃, EuBr₃ and EuI₃, respectively, whereas one obtains 6.16, 6.12, 6.09, and 6.08 at the HF level. Complete active space or multireference configuration interaction calculations lead to f populations of less than 6.1 electrons. The unphysical overfilling of the 4f shell leads for EuI₃ to a larger europium ion and thus also to a too long bond distance. The CCSD(T) bond distances of 2.027, 2.492, 2.650, and 2.836 Å for EuF₃, EuCl₃, EuBr₃, and EuI₃, respectively, compare favorably to the PBE0 results 2.027, 2.487, 2.650, and 2.919 Å only for the three lighter halogens. The same F–Eu–F bond angle of 113.8° was found for the pyramidal EuF₃ at the CCSD(T) and PBE0 levels, whereas for the three heavier halogens, a planar equilibrium geometry was found. Finally, the

Table 7. Ln–X Bond Distances (Å) of LnX₃ from AE DKH PBE0 Calculations Using the Extended Ln (23s16p12d6f)/[18s12p9d3f] Basis Sets Proposed in This Work, Relativistically Contracted F (11s6p2d1f)/[6s3p2d1f], Cl (14s9p2d1f)/[8s4p2d1f], Br (17s13p8d1f)/[10s8p4d1f], and I (19s15p11d2f)/[12s10p7d2f] Basis Sets for the Halides (Denoted TZVPP in the ORCA Basis Set Library), in Comparison to Averaged Recommended Values (rec.)^{64,63a}

Ln	LnF ₃		LnCl ₃		LnBr ₃		LnI ₃	
	this work	rec.	this work	rec.	this work	rec.	this work	rec.
La	2.106	2.108(31)	2.577	2.562(28)	2.731	2.715(26)	2.952	2.902(36)
Ce	2.084	2.098(29)	2.553	2.550(28)	2.705	2.700(23)	2.925	2.892(34)
Pr	2.068	2.075(17)	2.531	2.522(11)	2.684	2.679(14)	2.905	2.876(28)
Nd	2.065	2.069(20)	2.520	2.519(20)	2.670	2.669(17)	2.888	2.859(21)
Pm	2.051	2.058(19)	2.504	2.508(20)	2.656	2.656(17)	2.874	2.853(23)
Sm	2.036	2.047(18)	2.487	2.495(20)	2.646	2.644(17)	2.870	2.841(22)
Eu	2.027	2.037(17)	2.487	2.484(20)	2.650	2.633(17)	2.919	2.831(22)
Gd	2.027	2.031(22)	2.471	2.471(18)	2.619	2.622(18)	2.835	2.821(21)
Tb	2.018	2.015(14)	2.456	2.460(18)	2.609	2.608(18)	2.800	2.809(20)
Dy	2.003	2.005(14)	2.442	2.450(19)	2.591	2.596(19)	2.807	2.800(18)
Ho	1.999	1.994(13)	2.432	2.439(20)	2.583	2.585(19)	2.795	2.789(18)
Er	1.986	1.985(13)	2.421	2.427(20)	2.571	2.573(20)	2.783	2.779(17)
Tm	1.980	1.974(13)	2.410	2.417(20)	2.557	2.562(21)	2.773	2.769(17)
Yb	1.971	1.964(11)	2.399	2.404(20)	2.550	2.550(21)	2.768	2.758(16)
Lu	1.963	1.956(12)	2.390	2.395(22)	2.537	2.511(05)	2.752	2.752(19)
MAD	0.006		0.006		0.006		0.022	

^aThe numbers in parentheses denote the uncertainties for the last two digits as defined by the longer values of Hargittai and the shorter ones of Kovacs and Konings.

errors of the AE DKH PBE0 atomization energies with respect to the experimental values are always the highest for each series of lanthanide trihalides (*vide infra*), also confirming that Eu is a quite difficult case for DFT.

Finally, some brief comment concerning the structures applies. In our optimizations, we find the trifluorides to be slightly pyramidal (C_{3v}) and the other heavier trihalides to be essentially planar (D_{3h}), when the molecular point group is restricted to C_{3v} . Note that no symmetry constraints were imposed on the orbitals; i.e., for a single determinant description, symmetry breaking may occur. In particular, we do not find planar PmF_3 and SmF_3 molecules in variance with the results of Pantazis and Neese.¹² These authors comment that their observed structures are in remarkable agreement with the so-called asphericity model.⁶⁶ According to this model, originating from early work in solid state physics,⁶⁷ one has for LnX₃ pyramidal (Ln = Ce, Pr, Nd, Eu, Tb, Dy, Ho) and planar (Ln = La, Pm, Sm, Gd, Er, Tm, Yb, Lu) structures depending on the shape of the 4f shell charge distribution. It is noteworthy that already LaF₃, which was left out by Pantazis and Neese, does not obey this model. In addition, GdF₃ was found by these authors to be pyramidal and HoF₃ to be planar in disagreement with the model. Moreover, in view of the small energy differences between the pyramidal and planar structures, e.g., 0.036 eV for PmF_3 according to our PBE0 results, one may doubt if a single reference DFT model without consideration of SO coupling is able to describe accurately enough the LnX₃ electronic structure. In particular, it has to be noted that like most density functionals the PBE0 functional applied in this work as well as by Pantazis and Neese is not able to describe the degeneracy of different components of open shell states correctly, when these produce differing density distributions. This problem was already discussed, e.g., by Baerends and co-workers for main group and transition metals.²⁹ Already for the simplest case, the

Table 8. Mean Absolute Errors in LnX₃ PBE0 Atomization Energies (eV) of Scalar-Relativistic (SR) and Spin-Orbit Corrected (SO) Calculated Values with Respect to Experimental Data⁶⁸

method	LnF ₃		LnCl ₃		LnBr ₃		LnI ₃	
	SR	SO	SR	SO	SR	SO	SR	SO
AE DKH ^a	0.45	0.47	0.49	0.57	0.25	0.59	0.92	0.30
PP+CPP ^b	0.65	0.61	0.44	0.35	0.84	0.46	1.33	0.47
AE DKH ^c	0.47	0.47	0.32	0.28	0.45	0.27	0.81	0.24
AE DKH ^d	0.44	0.44	0.41	0.41	0.43	0.20	0.80	0.19
AE DKH ^e	0.49	0.52	0.44	0.45	0.42	0.27	0.75	0.24

^a Ref 12; SO corrections of this work were added. ^b 4f-in-core PP+CPP results.¹⁸ ^c This work using the proposed Ln basis sets and X TZVPP basis sets, MOLPRO. ^d This work using the Ln basis sets of ref 12 and X TZVPP basis sets, ORCA. ^e This work using the proposed Ln basis sets and X TZVPP basis sets, ORCA.

boron atom, degeneracy errors of up to 0.2 eV were found for the $2p^1 2P$ ground state at the GGA level. In relativistic AE DKS calculations, a degeneracy error of up to 0.6 eV was found for the $Yb^{3+} 4f^{13} 2F_{7/2}$ state at the LDA and GGA levels.³⁰ This defect resulted in a wrong ordering of molecular states for YbO. It is likely that such artifacts are also present for the LnX₃ open shell molecules and influence, besides the atomization energies, also the ground state equilibrium structures. We therefore do not discuss the bond angles here in detail and refer the reader interested in these results to the Supporting Information.

3.2.2. Atomization Energies. Table 8 summarizes the errors for the calculated LnX₃ atomization energies with respect to the data provided by Myers.⁶⁸ This author evaluated LnX₃ atomization

energies rounded to 1 kcal/mol (~ 0.04 eV) by using a thermodynamic cycle. He estimated the data to be accurate within ± 5 kcal/mol (± 0.22 eV) if only experimental values enter and ± 10 kcal/mol (± 0.44 eV) if one value had to be estimated. In view of these uncertainties, we neglected zero-point vibration corrections, which only amount to about 0.10, 0.06, 0.04, and 0.03 eV for LnF_3 , LnCl_3 , LnBr_3 , and LnI_3 , respectively.²⁶ Note that in the present calculations, no symmetry restrictions were imposed, and the lowest-energy KS determinant was searched for by applying rotations between occupied and empty 4f orbitals. In order to have an unbiased comparison, we performed analogous calculations with the basis sets of Pantazis and Neese and the ones proposed here. From the data for the heavier trihalides, i.e., LnBr_3 and LnI_3 , it is obvious that SO corrections improve the results significantly. It further can be seen that the overall quality of the results obtained with the proposed basis sets is similar to the one for the SARC basis sets.

It has to be noted, however, that since the applied DFT does not yield exact results, good performance may also partly be due to error compensations. Assuming a completely ionic electronic structure $\text{Ln}^{3+}(\text{F}^-)_3$ with a $4f^n$ ($n = 0-14$ for La–Lu) sub-configuration on the lanthanide, one may expect that errors in the first three lanthanide ionization potentials as well as the halogen electron affinity directly affect the results. Aside from the performance of the PBE0 functional, the basis set errors occurring, e.g., at the HF level, have to be considered. Whereas the errors in the halogen electron affinities affect equally the results for the present and the SARC basis sets, those in the sum of the first three lanthanide ionization potentials might lead to deviations. We note that the SARC basis sets yield somewhat larger errors (MAD 0.15 eV, maximum deviation 0.28 eV) than the present basis sets (MAD 0.03 eV, maximum deviation 0.08 eV).

Table 9 compares the DKH2 PBE0 LnX_3 atomization energies obtained with the proposed basis sets to the values obtained with the SARC basis of Pantazis and Neese.¹² Both sets produce very similar trends along the lanthanide series with a MAD of about 0.1 eV.

SO contributions to the atomization energies have been neglected in most of the previous work. The atomic and molecular SO total energy lowerings, however, can be substantial and usually do not cancel, as is obvious from the data compiled in Table 10. The molecular SO total energy contributions are very similar to the Ln^{3+} atomic energy lowerings, thus supporting models based on an ionic $\text{Ln}^{3+}(\text{X}^-)_3$ charge distribution. The results compiled in Table 10 are in disagreement with corresponding data derived by Adamo and Maldivi⁴² for LnX_3 ($\text{Ln} = \text{La}, \text{Lu}$) from AE DFT calculations using the ADF package. These authors report strong stabilizations of LaF_3 and LuF_3 by 1.23 and 1.17 eV, respectively, due to SO contributions. It is not clear how the closed-shell molecules can be stabilized with respect to the open-shell separated atoms. On the other hand, for LaI_3 and LuI_3 , a destabilization by 0.35 and 0.90 eV, respectively, due to SO effects was reported, which is at least in qualitative agreement with the present results. Still, it is unclear why the magnitude of the destabilization is so different for the two molecules, where the major contribution should arise from the three I atoms, whereas the much smaller contributions of the Ln atoms agree within 0.1 eV.

Finally, the basis set superposition error (BSSE) has to be considered. Corresponding corrections can be estimated according to the Boys and Bernardi counterpoise scheme⁶⁹ by calculating the atomic energies used for the evaluation of the atomization

Table 9. LnX_3 Atomization Energies (eV) from Spin-Free AE DKH2 PBE0 Calculations Using the Ln (23s16p12d6f)/-[18s12p9d3f] Basis Sets Proposed in This Work As Well As the Corresponding SARC Basis Sets¹² Together with F (11s6p1d)/[6s3p1d], Cl (14s9p1d)/[8s4p1d], Br (17s13p7d)/[10s8p3d], and I (19s15p10d)/[12s10p6d] Basis Sets (Denoted by TZVP in the ORCA Basis Set Library) for the Halides

Ln	LnF_3		LnCl_3		LnBr_3		LnI_3	
	this work	SARC	this work	SARC	this work	SARC	this work	SARC
La	19.86	19.71	15.78	15.66	14.30	14.17	12.43	12.32
Ce	19.77	19.80	15.67	15.70	14.20	14.22	12.31	12.34
Pr	18.76	19.05	14.66	14.96	13.18	13.45	11.28	11.57
Nd	18.09	18.32	14.08	14.30	12.58	12.80	10.70	10.92
Pm	17.75	17.93	13.71	13.89	12.22	12.40	10.35	10.53
Sm	16.64	16.80	12.62	12.78	11.14	11.31	9.32	9.48
Eu	15.66	15.79	11.63	11.77	10.19	10.33	8.53	8.65
Gd	19.16	19.16	15.20	15.22	13.69	13.71	11.81	11.83
Tb	18.89	18.91	14.89	14.92	13.38	13.41	11.49	11.52
Dy	18.04	17.99	14.02	14.00	12.51	12.47	10.60	10.58
Ho	17.74	17.67	13.75	13.72	12.23	12.18	10.33	10.30
Er	17.79	17.70	13.78	13.74	12.27	12.21	10.37	10.33
Tm	16.83	16.73	12.84	12.77	11.33	11.25	9.40	9.37
Yb	15.95	15.83	11.94	11.87	10.42	10.34	8.53	8.48
Lu	19.30	19.28	15.28	15.37	13.77	13.81	11.85	11.92
MAD	0.11		0.10		0.10		0.09	

Table 10. Atomic and Molecular Total Energy Lowerings (eV) Due to Spin-Orbit Interaction As Calculated from AE DKH2 Calculations Applying the Breit–Pauli–Hamiltonian in First-Order Perturbation Theory Based on Hartree–Fock Results

Ln	Ln^{3+}	LnF_3	LnCl_3	LnBr_3	LnI_3
La	-0.11	0.00	0.00	0.00	0.00
Ce	-0.08	-0.16	-0.14	-0.15	-0.15
Pr	-0.29	-0.28	-0.27	-0.28	-0.28
Nd	-0.34	-0.38	-0.36	-0.37	-0.37
Pm	-0.34	-0.44	-0.41	-0.43	-0.43
Sm	-0.26	-0.41	-0.39	-0.40	-0.40
Eu	0.00	-0.32	-0.23	-0.26	-0.27
Gd	-0.15	0.00	0.00	0.00	0.00
Tb	-0.52	-0.34	-0.31	-0.32	-0.29
Dy	-0.69	-0.62	-0.61	-0.61	-0.61
Ho	-0.76	-0.82	-0.81	-0.81	-0.81
Er	-0.70	-0.89	-0.89	-0.89	-0.89
Tm	-0.46	-0.81	-0.81	-0.81	-0.81
Yb	0.00	-0.51	-0.52	-0.53	-0.53
Lu	-0.20	0.00	0.00	0.00	0.00
		F	Cl	Br	I
		-0.02	-0.04	-0.15	-0.32

energies in the atomic and molecular basis sets. Taking the ytterbium trihalides as examples, one obtains with the MOLPRO code at the PBE0 level for the Yb basis set proposed here reductions of the atomization energies of 0.10, 0.09, 0.10, and

0.11 eV for YbF₃, YbCl₃, YbBr₃, and YbI₃, respectively, when TZVP basis sets are applied for the halogen atoms and all calculations are performed without symmetry restrictions using the molecular grid. The corresponding corrections are slightly smaller for the Yb SARC basis set of Pantazis and Neese; i.e., they amount to 0.07, 0.05, 0.05, and 0.07 eV, respectively. The corrections for the basis sets proposed here decrease upon the addition of a diffuse f function; e.g., for YbF₃, a value of 0.05 eV is obtained, whereas the addition of diffuse s, p, or d functions does not lead to significant changes.

An alternative to the evaluation of the atomization energies with respect to the neutral atoms is a corresponding evaluation with respect to the ions Ln³⁺ and X⁻. Taking YbBr₃ as an example, one obtains counterpoise corrections of 0.26 and 0.47 eV for the basis sets proposed here and the SARC sets of Pantazis and Neese, respectively. These large values almost exclusively arise from the contributions of the Br⁻ ions, whereas the Yb³⁺ ion contributes at most 0.01 eV for both lanthanide basis sets. The larger BSSE found for the SARC basis sets is due to the presence of the more diffuse functions than in the basis sets presented here.

In comparison to other errors of the calculations associated with the use of DFT, as well as the uncertainties of the experimental data, the relatively small BSSE found in the atomization energies calculated with respect to the neutral atoms is tolerable. For example, the atomic reference energies of the neutral halogen atoms in the atomic basis set were obtained by restricting the basis sets to the angular symmetries of the occupied shells, i.e., sp for F and Cl and spd for Br and I. By including the higher angular momentum functions used in the molecular calculations, i.e., d for F and Cl as well as f for Br and I, and allowing for symmetry breaking, lower total atomic energies are obtained. The energy lowerings by 0.07, 0.09, 0.04, and 0.05 eV for F, Cl, Br, and I, respectively, due to symmetry breaking are also automatically included in the atomic calculations performed in the molecular basis. Such contributions to the atomization energies arising from all three halogen atoms in a lanthanide trihalide are clearly larger than the BSSE corrections discussed above, which were therefore not evaluated for the final results compiled in Table 11. Note that corresponding energy lowerings would also arise for Ln atoms with open 4f and/or 5d shells.

The mean absolute deviations of the AE DKH PBE0 results from the experimental values of the atomization energies are largest for LnF₃ (0.47 eV) and very similar and about only half as large for the other trihalides LnCl₃ (0.28 eV), LnBr₃ (0.27 eV), and LnI₃ (0.24 eV), cf. Table 11. The mean unsigned relative errors are almost identical, i.e., 2.7, 2.0, 2.3, and 2.5% for LnF₃, LnCl₃, LnBr₃, and LnI₃, respectively. It is interesting to see that the mean absolute deviation of the four LnX₃ (X = F, Cl, Br, I) systems is for Eu (0.85 eV) more than twice as large as for any other lanthanide element Ln. The lowest mean absolute deviations occur for Gd (0.06 eV) and La (0.09 eV), whereas the values for the other elements range between 0.19 and 0.38 eV. This again might indicate that the description of Eu compounds at the DFT level is relatively poor, which was already obvious from the Eu–I bond distance (*vide supra*). In LaX₃, CeX₃, GdX₃, and LuX₃, the 4f occupation in the neutral atom is identical to the one in the trihalide molecule, i.e., the differential correlation contributions in the atomization energies should be small. In the case of Ce (0.27 eV), some problems may arise from the fact that the actual 4f¹5d¹6s²1G₄ ground state cannot be described at the DFT level and the higher ³F state had to be chosen as a reference,

Table 11. LnX₃ Atomization Energies (eV) from AE DKH PBE0 Calculations Using the Extended Ln (23s16p12d6f)/[18s12p9d3f] Basis Sets Proposed in This Work, Relativistically Contracted F (11s6p2d1f)/[6s3p2d1f], Cl (14s9p2d1f)/[8s4p2d1f], Br (17s13p8d1f)/[10s8p4d1f], and I (19s15p11d2f)/[12s10p7d2f] Basis Sets for the Halides (Denoted TZVPP in the ORCA Basis Set Library), in Comparison to Experimental Values (exptl.)^{68a}

Ln	LnF ₃		LnCl ₃		LnBr ₃		LnI ₃	
	this work	exptl.	this work	exptl.	this work	exptl.	this work	exptl.
La	19.94	19.86	15.83	15.87	13.80	13.70	11.41	11.28
Ce	20.09	20.08	15.98	15.57	13.93	13.57	11.54	11.19
Pr	19.00	19.08	14.87	15.22	12.82	13.18	10.42	10.58
Nd	18.30	19.04	14.26	14.48	12.22	12.62	9.80	9.93
Pm	18.12		14.08		12.03		9.64	
Sm	16.97	17.30	12.91	13.23	10.92	11.10	8.57	9.06
Eu	16.09	17.22	12.06	13.05	10.08	10.80	7.89	8.46
Gd	19.19	19.21	15.22	15.13	13.16	13.05	10.76	10.80
Tb	18.86	18.99	14.87	15.13	12.80	12.88	10.37	10.67
Dy	18.01	17.35	13.92	14.18	11.85	11.84	9.44	9.67
Ho	17.98	17.22	14.06	14.31	11.98	11.75	9.57	9.37
Er	18.19	17.22	14.19	14.31	11.96	12.14	9.70	9.93
Tm	17.39	17.04	13.39	13.66	11.30	11.10	8.91	8.98
Yb	16.56	16.05	12.67	12.40	10.46	10.06	8.19	7.85
Lu	19.28	18.43	15.29	15.18	13.21	12.79	10.78	10.67
MAD	0.47		0.28		0.27		0.24	

^a Spin-orbit corrections have been applied to the theoretical results.

whereas for Lu the larger deviations (0.37 eV) cannot be explained. We note however that large-scale AE DKH CCSD(T) calculations yielded without and with the counterpoise correction of the BSSE an atomization energy of 20.75 and 19.79 eV, respectively.⁴¹ After adding SO corrections, the latter value reduces to 18.63 eV, which is much closer to the experimental value of 18.43 eV, i.e., about 18.53 eV after subtracting the zero-point vibration energy, compared to the SO-corrected AE DKH PBE0 result of 19.28 eV. These results for the relatively simple case of LuF₃ may point to some deficiencies of the PBE0 functional, which are not only present for open-4f-shell cases.

3.3. Conclusions. Segmented contracted lanthanide basis sets of modest size to be used in connection with the Douglas–Kroll–Hess second-order relativistic Hamiltonian and preferably density functional theory have been presented. The basis sets provide lower total energies than the previously proposed SARC basis sets of the same size and contraction pattern, and they give a more faithful description of electronic states with different valence occupations. Despite their slightly less diffuse outermost functions, they yield satisfactory results in molecular test calculations on the lanthanide trihalides. It was found that the one-center approximation for the evaluation of Douglas–Kroll–Hess second-order relativistic contributions leads on the average to 0.02 Å too short bond distances for the lanthanide triiodides. In addition, the single reference density functional theory description leads for some molecules to noticeable distortions from the ideal C_{3v} and D_{3h} geometries. Due to the uncertainties in the experimental reference data, a more rigorous calibration of the results of the present study has to wait for higher level ab

initio calculations using larger basis sets and also including spin-orbit effects.

■ ASSOCIATED CONTENT

S Supporting Information. Listings of segmented contracted basis sets as well as even-tempered basis sets used to generate the Hartree-Fock-limit reference data. Structural and energetic information on all 60 lanthanide trihalides obtained with various basis sets at the density functional theory level. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: m.dolg@uni-koeln.de.

■ ACKNOWLEDGMENT

Several comments of D. Pantazis and F. Neese on their previous work are gratefully acknowledged. The author thanks X. Cao for patiently waiting for the end of this project.

■ REFERENCES

- (1) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742–3748.
- (2) Jansen, G.; Hess, B. A. *Phys. Rev. A* **1989**, *39*, 5016–6017.
- (3) Wolf, A.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2002**, *117*, 9215–9226.
- (4) Nakajima, T.; Hirao, K. *Chem. Rev.* **2011**, [Online] DOI: 10.1021/cr200040s.
- (5) Tsuchiya, T.; Abe, M.; Nakajima, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 4463–4472.
- (6) Nakajima, T.; Hirao, K. *J. Chem. Phys.* **2002**, *116*, 8270–8275.
- (7) Sekiya, M.; Noro, T.; Miyoshi, E.; Osanai, Y.; Koga, T. *J. Comput. Chem.* **2006**, *27*, 463–470.
- (8) Sekiya, M.; Noro, T.; Koga, T.; Saito, S. L. *J. Comput. Chem.* **2010**, *31*, 497–499.
- (9) Almlöf, J.; Taylor, P. J. *Chem. Phys.* **1987**, *86*, 4070–4077.
- (10) Roos, B. O.; Lindh, R.; Malmqvist, P. A.; Veryazov, V.; Widmark, P. O.; Borin, A. C. *J. Phys. Chem. A* **2008**, *112*, 11431–11435.
- (11) Peterson, K. Gaussian Basis Sets for Quantum Mechanical (QM) Calculations. In *Computational Inorganic and Bioinorganic Chemistry*; Solomon, E. I., King, R. B., Scott, R. A., Eds.; Wiley: New York, 2009; pp 187–199.
- (12) Pantazis, D. A.; Neese, F. *J. Chem. Theory Comput.* **2009**, *6*, 2229–2238.
- (13) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6169.
- (14) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (15) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (16) Dolg, M.; Stoll, H.; Preuß, H. *J. Chem. Phys.* **1989**, *90*, 1730–1734.
- (17) Cao, X.; Dolg, M. *THEOCHEM* **2002**, *581*, 139–147.
- (18) Dolg, M. *AIP Conf. Proc.* **2011** In press.
- (19) Dolg, M.; Stoll, H. Electronic structure calculations for molecules containing lanthanide atoms. In *Handbook on the Physics and Chemistry of Rare Earths*; Gschneidner, K. A., Jr., Eyring, L., Eds.; Elsevier: Amsterdam, 1996; Vol. 22, pp 607–729.
- (20) Werner, H.-J.; Knowles, P. J. MOLPRO 2006.1. <http://www.molpro.net> (accessed August 19, 2011).
- (21) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053–5063.
- (22) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259–267.
- (23) Dolg, M. Fortran code for basis set and geometry optimizations. Unpublished.
- (24) Cao, X.; Dolg, M. *J. Chem. Phys.* **2001**, *115*, 7348–7355.
- (25) Cao, X.; Dolg, M. *THEOCHEM* **2002**, *581*, 139–147.
- (26) Dolg, M. In preparation, 2011.
- (27) Neese, F. ORCA 2.7–00. <http://www.thch.uni-bonn.de/tc/orca> (accessed August 19, 2011).
- (28) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. J. *Mol. Phys.* **2004**, *102*, 2311–2321.
- (29) Baerends, E. J.; Branchadell, V.; Sodupe, M. *Chem. Phys. Lett.* **1997**, *265*, 481–489.
- (30) Liu, W.; Dolg, M.; Li, L. *J. Chem. Phys.* **1998**, *108*, 2886–2895.
- (31) Pyykkö, P. *Adv. Quantum Chem.* **1978**, *11*, 353–409.
- (32) Pyykkö, P. *Chem. Rev.* **1988**, *88*, 563–594.
- (33) Dolg, M. *Mol. Phys.* **1996**, *88*, 1645–1655.
- (34) Berning, A.; Schweizer, M.; Werner, H.-J.; Knowles, P. J.; Palmieri, P. *Mol. Phys.* **2000**, *98*, 1823–1833.
- (35) Dolg, M.; Stoll, H.; Preuß, H. *THEOCHEM* **1991**, *235*, 67–79.
- (36) Cundari, T. R.; Stevens, W. J. *J. Chem. Phys.* **1993**, *98*, 5555–5565.
- (37) Joubert, L.; Picard, G.; Legendre, J.-J. *Inorg. Chem.* **1998**, *37*, 1984–1991.
- (38) Joubert, L.; Silvi, B.; Picard, G. *Theor. Chem. Acc.* **2000**, *104*, 109–115.
- (39) Lanza, G.; Varga, Z.; Kolonits, M.; Hargittai, M. *J. Chem. Phys.* **2008**, *128*, no. 074301.
- (40) Tsukamoto, S.; Mori, H.; Tatewaki, H.; Miyoshi, E. *Chem. Phys. Lett.* **2009**, *474*, 28–32.
- (41) Weigand, A.; Cao, X.; Yang, J.; Dolg, M. *Theor. Chem. Acc.* **2010**, *126*, 117–127.
- (42) Adamo, C.; Maldivi, P. *J. Phys. Chem. A* **1998**, *102*, 6812–6820.
- (43) Hampel, C.; Peterson, K.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1–12.
- (44) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219–5227.
- (45) Jorge, F. E.; de Castro, E. V. R.; da Silva, A. B. F. *J. Comput. Chem.* **1997**, *18*, 1565–1569.
- (46) de Castro, E. V. R.; Jorge, F. E. *J. Chem. Phys.* **1998**, *108*, 5225–5229.
- (47) Froese-Fischer, C. *The Hartree-Fock Method for Atoms*; Wiley: New York, 1977.
- (48) Dolg, M.; Cao, X. The relativistic energy-consistent ab initio pseudopotential approach and its application to lanthanide and actinide compounds. In *Recent Advances in Computational Chemistry*; Hirao, K., Ishikawa, Y., Eds.; World Scientific: River Edge, NJ, 2009; Vol. 6, pp 1–35.
- (49) Dolg, M.; Cao, X. Lanthanides and Actinides: Computational Methods. In *Computational Inorganic and Bioinorganic Chemistry*; Solomon, E. I., King, R. B., Scott, R. A., Eds.; Wiley: New York, 2009; pp 503–515.
- (50) National Institute of Standards, Atomic Spectra Database. <http://www.nist.gov/pml/data/asd.cfm> (accessed August 19, 2011).
- (51) Metz, B.; Schweizer, M.; Stoll, H.; Dolg, M.; Liu, W. *Theor. Chem. Acc.* **2000**, *104*, 22–28.
- (52) Metz, B.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2000**, *113*, 2563–2569.
- (53) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113–11123.
- (54) Figgen, D.; Rauhut, G.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2005**, *311*, 227–244.
- (55) Peterson, K. A.; Figgen, D.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2007**, *126*, no. 124101.
- (56) Figgen, D.; Peterson, K. A.; Dolg, M.; Stoll, H. *J. Chem. Phys.* **2009**, *130*, no. 164108.
- (57) Cao, X.; Dolg, M. *J. Phys. Chem. A* **2009**, *113*, 12573–12581.
- (58) Weigand, A.; Cao, X.; Vallet, V.; Flament, J.-P.; Dolg, M. *J. Phys. Chem. A* **2009**, *113*, 11509–11516.
- (59) Liu, W.; Dolg, M. *Phys. Rev. A* **1998**, *57*, 1721–1728.

- (60) Weigend, F. Personal communication, 2011.
- (61) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (62) Weigend, F.; Baldes, A. *Phys. Chem. Chem. Phys.* **2010**, *133*, no. 174102.
- (63) Kovacs, A.; Konings, R. J. M. *J. Phys. Chem. Ref. Data* **2004**, *33*, 377–404.
- (64) Hargittai, M. *Coord. Chem. Rev.* **1988**, *91*, 35–88.
- (65) Pantazis, D.; Neese, F. Personal communication, 2010.
- (66) Molnár, J.; Hargittai, M. *J. Phys. Chem.* **1995**, *99*, 10780–10784.
- (67) Sievers, J. *Z. Phys. B: Condens. Matter* **1982**, *45*, 289–296.
- (68) Myers, C. *Inorg. Chem.* **1975**, *14*, 199–201.
- (69) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules

Pengyu Ren,^{*,†} Chuanjie Wu,[‡] and Jay W. Ponder^{*,‡}

[†]Department of Biomedical Engineering, The University of Texas at Austin, Austin, Texas 78712, United States

[‡]Department of Chemistry and Department of Biochemistry & Molecular Biophysics, Washington University, St. Louis, Missouri 63130, United States

 Supporting Information

ABSTRACT: An empirical potential based on permanent atomic multipoles and atomic induced dipoles is reported for alkanes, alcohols, amines, sulfides, aldehydes, carboxylic acids, amides, aromatics, and other small organic molecules. Permanent atomic multipole moments through quadrupole moments have been derived from gas phase *ab initio* molecular orbital calculations. The van der Waals parameters are obtained by fitting to gas phase homodimer QM energies and structures, as well as experimental densities and heats of vaporization of neat liquids. As a validation, the hydrogen bonding energies and structures of gas phase heterodimers with water are evaluated using the resulting potential. For 32 homo- and heterodimers, the association energy agrees with *ab initio* results to within 0.4 kcal/mol. The RMS deviation of the hydrogen bond distance from QM optimized geometry is less than 0.06 Å. In addition, liquid self-diffusion and static dielectric constants computed from a molecular dynamics simulation are consistent with experimental values. The force field is also used to compute the solvation free energy of 27 compounds not included in the parametrization process, with a RMS error of 0.69 kcal/mol. The results obtained in this study suggest that the AMOEBA force field performs well across different environments and phases. The key algorithms involved in the electrostatic model and a protocol for developing parameters are detailed to facilitate extension to additional molecular systems.

INTRODUCTION

Organic molecules are the basic constituents of biology and material science. Modeling studies involving organic compounds are widely used in many areas such as physical chemistry, biological structure and function, and nanotechnology. Progress in quantum chemistry and the availability of fast computers has empowered the routine study of small molecules with high levels of *ab initio* theory and large basis sets. However, first principles statistical thermodynamics sampling techniques are still not practical for use with most high-level QM methods. Thus, molecular modeling based on empirical potentials is widely used for theoretical inquiries into microscopic and macroscopic phenomena across chemistry and biology. Atom-based force field models such as MM3,¹ AMBER,² CHARMM,³ OPLS,⁴ and GROMOS⁵ have been developed for a wide range of organic compounds and biomacromolecules. These models describe electrostatic interactions with fixed point charges on atoms and treat van der Waals interactions via Lennard-Jones potentials or other simple functions. Numerous studies have shown that many of the physical properties and structures of organic molecules can be adequately reproduced with current fixed charge force fields. Increases in computing power have enabled the simulation of larger molecular systems and more precise investigation of their properties. However, there are acknowledged shortcomings of the current generation of fixed charge potentials. They assume the atomic charges derived from training systems are approximately transferable to systems in different chemical environments. Explicit accounting of many-body effects is required for a general potential to capture the electrostatic response to different

molecular environments: homo- or heterogeneous, low or high dielectric, nonpolar or highly polarizable.

Polarization effects were initially used in the description of molecular refractivity and other chemical phenomena nearly 100 years ago.⁶ Early in the era of modern computational chemistry, polarization was applied to the study of enzymatic reactions⁷ and incorporated into prototype molecular dynamics algorithms.⁸ Recently, there have been increasing efforts toward developing polarizable force fields for molecular simulation, based on a variety of empirical models for induction, such as classical induced dipoles,^{2,9–22} fluctuating charges,^{23–30} and Drude oscillators.^{9,31–35} Detailed discussions of the various polarization models can be found in recent reviews of polarizable force field development.^{36–40} The performance of different approaches in accounting for polarization has been compared in the study of ion and small molecule interactions.^{41,42} The modeling of neat organic liquids, including alcohols, acids, amides, and aromatics, has also been reported using polarizable potentials.^{11,22,35,43–50}

Restriction to fixed atomic point charges constrains the flexibility of a model in representing the electrostatic potential around a molecule^{51,52} and thus limits the accuracy of the treatment of molecular interactions. Improvement can be achieved by adding extra charge sites, typically at bond centers or lone pair positions. For example, the TIPxP series of water models, TIP3P,⁵³ TIP4P,⁵³ and TIP5P,⁵⁴ adopts increasing numbers of charge sites. Recently, the extra site approach was introduced into a Drude

Received: May 2, 2011

Published: August 31, 2011

oscillator-based polarizable model as a way to address the anisotropy in atomic charge distribution due to lone pair electrons.⁵⁰ Alternatively, one can directly incorporate higher order moments, such as dipole and quadrupole moments, at the atomic centers to improve the representation of the charge distribution. The convergence advantage of using multipoles distributed over atomic sites, as opposed to a single molecule-centered set of moments, has been discussed in the literature.^{55,56} Over two decades ago, Buckingham and Fowler^{57,58} were the first to apply distributed multipole moments to structural modeling of small molecule complexes. Their proposed intermolecular potential, consisting of hard sphere repulsion and atomic multipole-based electrostatics, was able to reproduce a number of experimental equilibrium geometries and orientational preferences. Recently, coarse-grained potentials with point multipoles have been used successfully in modeling hydrogen-bonded molecular liquids.^{59,60}

The AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field was initially developed for water.^{18,20} The current study reports the extension of the AMOEBA model to organic compounds including alkanes, alcohols, amines, sulfides, aldehydes, carboxylic acids, amides, and aromatics. A cornerstone of the AMOEBA force field is an improved electrostatic potential based on atomic multipoles and classical induced dipole moments. The atomic multipole moments are obtained from high-level *ab initio* calculations on gas phase monomers. An empirical atomic dipole induction model describes the many-body polarization effects important in clusters and condensed phase environments. A small, consistent set of atomic polarizability parameters is used to treat intermolecular polarization as well as intramolecular polarization between functional group fragments. van der Waals (vdW) parameters are refined via gas-phase homodimer molecular orbital calculations and molecular dynamics simulation of liquid properties. Additional gas-phase and liquid-phase computations, including hydrogen bonding in gas-phase heterodimers, the dielectric and diffusion constants of neat liquids, and hydration free energies of organic compounds have been utilized to validate the resulting force field.

METHODS

Potential Energy Model. The interaction energy among atoms is expressed as

$$U = U_{\text{bond}} + U_{\text{angle}} + U_{\text{b-a}} + U_{\text{oop}} + U_{\text{torsion}} + U_{\text{vdW}} + U_{\text{ele}}^{\text{perm}} + U_{\text{ele}}^{\text{ind}} \quad (1)$$

where the first five terms describe the short-range valence interactions: bond stretching, angle bending, bond–angle cross term, out-of-plane bending, and torsional rotation. The last three terms are the nonbonded interactions: van der Waals, permanent electrostatic, and induced electrostatic contributions. The individual terms for these interactions have been described in detail in a previous publication.⁶¹ Some additional methodology, introduced to treat electrostatic polarization in molecular systems beyond water, will be detailed below. Polarization effects in AMOEBA are treated via Thole's interactive induction model that utilizes distributed atomic polarizability.^{62,63} According to this interactive induction scheme, induced dipoles produced at the atomic centers mutually polarize all other sites. A damping function is used at short-range to eliminate the polarization

catastrophe and results in correct anisotropy of the molecular response (i.e., diagonal components of the molecular polarizability tensor) starting from isotropic atomic polarizabilities. Thole damping is achieved by screening of pairwise atomic multipole interactions and is equivalent to replacing a point multipole moment with a smeared charge distribution.¹³ The damping function for charges is given by

$$\rho = \frac{3a}{4\pi} \exp(-au^3) \quad (2)$$

where $u = r_{ij}/(\alpha_i\alpha_j)^{1/6}$ is the effective distance as a function of interatomic distance r_{ij} and the atomic polarizabilities of atoms i (α_i) and j (α_j). The coefficient a is the dimensionless width of the smeared charge distribution and controls the damping strength. The corresponding damping functions for charge, dipole, and quadrupole interactions were reported previously.¹⁸

The Thole model is able to reproduce the molecular polarizability tensors of numerous small molecules with reasonable accuracy using only element-based isotropic atomic polarizabilities and a single value for the damping factor.⁶² In our water study, it was discovered that the dependence of molecular polarizability on the damping coefficient is weak, but the polarization energy is much more sensitive to the strength of damping. After fitting the interaction energies of a series of small water clusters, we have chosen a universal damping factor of $a = 0.39$, rather than the value of 0.572 suggested by Thole. We adopt the atomic polarizabilities (\AA^3) as originally given by Thole, i.e., 1.334 for carbon, 0.496 for hydrogen, 1.073 for nitrogen, and 0.837 for oxygen. The only exception is for aromatic carbon and hydrogen atoms, where we found that the use of larger values greatly improves the molecular polarizability tensor of benzene and polycyclic aromatics. The AMOEBA values for atomic polarizability are given in Table 1. In addition, for metal dications, we have found it necessary to use stronger damping ($a < 0.39$) to better represent the electric field around the ions.^{21,61,64}

Intramolecular Polarization. For a large molecule such as a multifunctional organic or a biopolymer, polarization arises not only from the electric field of other molecules but also from distal portions of the same molecule. It is crucial to describe the intra- and the intermolecular response in a consistent manner. In prior work, we investigated the effect of intramolecular polarization on the conformational dependence of the electrostatic potential surrounding a dipeptide.¹⁵ As observed by others,⁶⁵ the electrostatic parameters derived for alanine dipeptide vary significantly depending upon the conformation used to derive the values. A simple average over multipole moments obtained from a set of conformers does not transfer well between conformers, i.e., gives poor electrostatic potentials on different conformers. Furthermore, when short-range polarization between bonded atoms is ignored, use of intramolecular polarization yields only marginal improvement over current nonpolarizable potentials. To overcome this problem, a group based intramolecular polarization scheme has been devised.¹⁵ The “groups” are typically functional groups with limited conformational degrees of freedom, such as an amide group or phenyl ring. *In this scheme, the permanent atomic multipoles (PAM) polarize between, and not within, groups. For a small molecule consisting of a single polarization group, such as water or ammonia, permanent atomic multipoles do not polarize sites within the same molecule, while “mutual” induction occurs among all polarizable sites as described above.* This design offers a clear connection between the treatment of small molecules and that of the analogous fragments inside a larger

Table 1. vdW Parameters and Atomic Polarizabilities for AMOEBA Atom Classes

atom	description	R^0 (Å)	ϵ (kcal/mol)	polarizability (Å ³)
C	alkane (CH ₃ – or –CH ₂ –)	3.820	0.101	1.334
H	alkane (CH ₃ –)	2.960	0.024 (0.92)	0.496
H	alkane (–CH ₂ –)	2.980	0.024 (0.94)	0.496
C	alkane (–CH<)	3.650	0.101	1.334
H	alkane (–CH<)	2.980	0.024 (0.94)	0.496
O	hydroxyl (water, alcohol)	3.405	0.110	0.837
H	hydroxyl (water, alcohol)	2.665	0.0135 (0.91)	0.496
O	carbonyl (aldehyde, amide, acid)	3.300	0.112	0.837
H	acid (HO)	2.665	0.0150 (0.91)	0.496
C	carbonyl (aldehyde, amide, acid)	3.820	0.106	1.334
C	aromatic carbon	3.800	0.091	1.750
H	aromatic (HC)	2.980	0.026 (0.92)	0.696
N	amine nitrogen (ammonia, amine)	3.710	0.105	1.073
H	amine (HN)	2.700	0.020 (0.91)	0.496
N	amide nitrogen	3.710	0.110	1.073
H	amide (HN)	2.590	0.022 (0.90)	0.496
S	sulfur	4.005	0.355	2.800
H	sulfhydryl (HS)	2.770	0.024 (0.96)	0.496

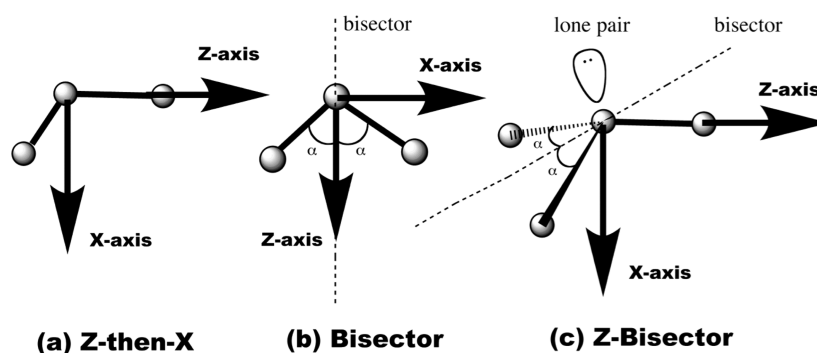


Figure 1. Local coordinate frame definitions for atomic multipole sites. (a) The Z-then-X frame is used for general sites and, with the addition of a third orthogonal y axis, can treat chiral centers. The majority of AMOEBA multipole sites are defined using this local frame. (b) The Bisector frame is useful for molecules with 2-fold local symmetry or pseudosymmetry, such as water and aliphatic methylene carbon atoms. (c) The Z-Bisector frame is used for sites such as the sulfur atom of dimethylsulfoxide, which have a distinct primary (“Z”) axis and symmetry or pseudosymmetry along a secondary direction.

molecule, thereby facilitating the transfer of PAM values from small model compounds to larger species such as polypeptides and nucleic acids.

We use Distributed Multipole Analysis⁶⁶ (DMA) to extract atomic multipoles from *ab initio* calculations. Starting from the DMA atomic multipoles for an arbitrary conformer of a model compound, M_i^{DMA} , one can derive the intrinsic “permanent” atomic multipole moments, M_i , that satisfy

$$M_i^{\text{DMA}} = M_i + \mu_i \quad (3)$$

where μ_i is the dipole induced by intramolecular polarization by M_i . The M_i are obtained by substituting μ_i from eq 4 below into the preceding equation. This approach allows derivation of conformation-independent atomic multipole parameters for larger organic compounds with multiple polarization groups. A local coordinate frame is defined at each atomic site and used to rotate atomic multipole moments as neighboring atoms move during structure manipulation. As shown in Figure 1, three types

of local frames are sufficient to handle essentially all situations arising in organic chemistry.

Polarization Energy. Formally, the induced dipole vector on any polarizable site i can be expressed as

$$\mu_i^{\text{ind}} = \alpha_i \left(\sum_{j \neq i} T_{ij}^1 M_j + \sum_{k \neq i} T_{ik}^{11} M_k \right) \quad (4)$$

and the associated energy is

$$U_{\text{ele}}^{\text{ind}} = -\frac{1}{2} \sum_i (\mu_i^{\text{ind}})^T E_i \quad (5)$$

where $T_{ij}^1[\nabla^1, \nabla^2, \nabla^3, \text{etc.}]$ is a 3×13 matrix with $\nabla^{l+m+n} = \partial^l/\partial x^l \partial^m/\partial y^m \partial^n/\partial z^n$ representing the second through fourth rows of the multipole–multipole interaction matrix T_{ij} (see Appendix, eq A2). $T_{ij}^{11} = \nabla_{ik}^2$ is a 3×3 submatrix consisting of elements in T_{ij}^1 corresponding to the dipole moments. As discussed above, the atomic polarizability is isotropic. Therefore, the off-diagonal elements of the tensor, α_{ij} , are all zero, and the

three diagonal elements take the same scalar value. The factor of 1/2 is a result of the induction cost for the formation of induced dipoles.

In eq 4, the first term inside the parentheses on the right-hand side is the “direct” electric field, E , due to permanent multipoles outside the polarization group of atom i (index j). The second term corresponds to “mutual” induction by all induced dipoles (index k). Thus, direct induction due to permanent multipoles only occurs between groups, while mutual polarization between induced dipoles involves every atom pair. When computing energies, as opposed to induced dipoles, the scaling of 1–2, 1–3, and other local interactions is applied to permanent and polarized electrostatic terms as with other molecular mechanics models. In addition, dipole induction is damped at short-range to avoid the “polarization catastrophe”, and damping is applied consistently to the induced field, energy, and force. For convenience, scaling and damping is assumed to be implicitly included with the T matrix elements in the present discussion.

The set of induced dipole equations is solved iteratively to obtain the final dipole values. The convergence is accelerated via a successive over-relaxation (SOR) procedure.⁶⁷

$$\mu_i(n+1) = (1-\omega)\mu_i(n) + \omega[\mu_i(0) + \alpha_i \sum_{\{k\}} T_{ik}^{11} \mu_k(n)] \quad (6)$$

where $\mu_i(0) = \alpha_i \sum_j T_{ij}^1 M_j$ is the “direct” induced dipole moment generated by the permanent field. The default ω value is 0.7, while for the case $\omega = 1$, eq 6 reduces to eq 4.

Energy Gradient and Ewald Summation. The energy gradient due to permanent multipole moments, including force and torque components, was derived by Smith for a standard Ewald summation.⁶⁸ We have previously reported the AMOEBA Ewald force, torque, and virial arising from dipole induction in water systems.¹⁸ Note that the pairwise direct (non-Ewald) formula can be obtained by replacing the real-space screening factor $B(r)$ with the corresponding function of $1/r$ ^{18,68} and vice versa. The torque components are converted to atomic forces on the relevant frame-defining atoms in our implementation. It is also possible to derive the analytical forces corresponding to the torques directly via an infinitesimal rotation,⁶⁹ or by taking the derivative of the rotation matrix.⁷⁰ When evaluating the energy derivative directly, the additional chain rule terms due to the local frame rotation matrices are equivalent to the forces converted by means of the torque implementation. In the Appendix, we provide a derivation of the polarization energy gradient, with a focus on terms arising from intramolecular polarization.

The Ewald real-space interactions need to be modified to accommodate short-range scaling of electrostatics and damping of dipole induction as mentioned above. To scale the interaction between an atom pair, a term $(f_{\text{scale}} - 1) U'$ is added to the total Ewald energy, where U' is the full (non-Ewald) interaction between the pair, and the scaling factor, f_{scale} , ranges from 0 to 1. Analogous approaches are used in computing forces, fields, and torques.

Particle-mesh Ewald (PME) for point multipoles⁶⁹ has been implemented in the TINKER and AMBER/PMEMD software packages. PME significantly improves the computational efficiency as its cost scales as $N \log N$, where N is the number of particles. The addition of dipole and quadrupole moments to the PME method roughly doubles the computational expense versus point charge only models. Calculation of induced dipoles can be

time-consuming with the simple iterative solution method, depending upon the level of SCF convergence required. Alternative fast predictive induced dipole schemes have been suggested.^{71,72} Acceleration via extended Lagrangian methods has been reported for induced dipole polarization^{73–75} and is under investigation for the AMOEBA model.

A standard Ewald summation implies the use of “tin foil” boundary conditions, corresponding to a system immersed in a conducting dielectric environment (i.e., $\epsilon = \infty$). It is possible to include a boundary correction to the Ewald energy if other environments, such as insulating boundary conditions, are desired. For a cubic box, the correction term is a function of the total cell dipole moment, while for other system shapes the analytical form is difficult to derive.^{76,77} Note that the energy obtained via Ewald summation is equivalent to the energy obtained using an infinitely long atom-based cutoff for the same periodic system. However, group-based cutoffs are often applied to preserve local charge neutrality. When using group cutoffs, the energy asymptotically approaches a different value from atom cutoffs as the cutoff length increases. The difference between the two energies is exactly equal to the above boundary correction term. This suggests that care must be taken if cutoff methods are applied to a system containing multipoles since the dipole and higher order moments are intrinsically group based.

Parameterization. The atomic polarizabilities are listed for each AMOEBA atom type in Table 1. The values are the same as those derived by Thole⁶² except for aromatic carbon and hydrogen atoms, which have been systematically refined using a series of aromatic systems, including a small carbon nanotube (see Table 3). The molecular polarizabilities computed using the current model are compared to experimental values for selected compounds in Table 2. Reducing the damping factor from Thole’s original value of 0.567 to AMOEBA’s 0.39 is critical to correctly reproducing water cluster energetics.¹⁸ On the other hand, AMOEBA’s greater damping leads to a slight systematic underestimation of molecular polarizabilities. However, given the simplicity of the model, the agreement is generally satisfactory for both average polarizabilities and their anisotropies. As described above, polarization groups are defined for purposes of treating intramolecular polarization. Typically, a functional group is treated as a single polarization group. For example, methylamine is a group by itself, while ethylamine has two groups: $-\text{CH}_2\text{NH}_2$ and CH_3- . The groups are specified in AMOEBA parameter files in the following format: “polarize A α 0.39 B C”, where α is the polarizability for atom type A, 0.39 is the damping coefficient in eq 2, and B and C are possible bonded atom types that belong to the same polarization group as atom type A.

The permanent atomic multipoles were derived for each molecule from *ab initio* QM calculations. *Ab initio* geometry optimization and a subsequent single-point energy evaluation were performed at the MP2/6-311G(1d,1p) level using Gaussian 03.⁷⁸ For small molecules with less than six heavy atoms, Distributed Multipole Analysis (DMA v1.2⁷⁹) was used to compute the atomic multipole moments in the global frame using the density matrix from the QM calculation. Next, the TINKER POLEDIT program rotates the atomic multipoles into a local frame and extracts Thole-based intramolecular polarization to produce permanent atomic multipole (PAM) parameters. Thus, when the AMOEBA polarization model is applied to the permanent atomic moments, the original *ab initio* derived DMA is recovered. Finally, the POTENTIAL program from the TINKER package is used to optimize the permanent atomic multipole




Table 2. Comparison of Experimental and Computed Molecular Polarizabilities (\AA^3)^a

		α_{avg}	α_x	α_y	α_z
methane	AMOEBA	2.48	2.48	2.48	2.48
	Thole	2.55	2.55	2.55	2.55
	exptl	2.62	2.62	2.62	2.62
ethane		4.25	4.66	4.05	4.05
		4.46	4.93	4.24	4.24
		4.48	4.99	4.22	4.22
propane		6.01	6.75	5.78	5.51
		6.29	7.18	5.98	5.68
		6.38	7.66	5.74	5.74
formaldehyde		2.44	2.77	2.55	2.01
		2.54	3.07	2.70	1.86
		2.45	2.76	2.76	1.83
formamide		3.65	4.32	3.87	2.74
		3.79	4.86	4.04	2.50
		4.08 (4.22)	5.24	$(\alpha_y + \alpha_z = 7.01)$	
acetamide		5.43	6.26	5.72	4.30
		5.71	6.70	6.30	4.13
		5.67	6.70	$(\alpha_y + \alpha_z = 10.3)$	
methanol		3.19	3.61	3.02	2.93
		3.35	3.92	3.13	2.99
		3.32 (3.26)	4.09	3.23	2.65
ethanol		4.94	5.44	4.84	4.54
		5.08	5.76	4.98	4.50
		5.26 (5.13)	6.39	4.82	4.55
propanol		6.73	7.63	6.53	6.03
		7.21	8.42	6.89	6.30
		6.97 (6.96)			
NH ₃		1.92	2.07	2.07	1.62
		1.95	2.17	2.17	1.52
		2.22			
dimethylether		4.99	5.92	4.55	4.52
		5.24	6.55	4.58	4.57
		5.24	6.38	4.94	4.39
benzene		9.68	11.42	11.42	6.20
		9.71	11.70	11.70	5.72
		9.01 (10.44)	11.03	11.03	4.97

^a Experimental data are taken from Tables 5 and 6 of Applequist et al.¹⁴⁶ Where available, more recent experimental values for α_{avg} from Bosque and Sales¹⁴⁷ are reported in parentheses.

parameters by fitting to the electrostatic potential on a grid of points outside the vdW envelope of the molecule. The reference potential for the fitting step is typically derived from a single point calculation at the MP2/aug-cc-pVTZ level. Only a partial optimization to the potential grid is used to keep the atomic moments close to their DMA-derived values while still providing an improved molecular potential. The fitting approach is also useful for molecules containing symmetry-averaged atoms of the same atomic multipole type. In this case, simple arithmetic averaging would degrade the quality of the PAM. For example, in dimethyl- or trimethylamine, all of the methyl hydrogen atoms are indistinguishable and adopt the same atom type. The DMA multipole values for these atoms are somewhat different due their nonequivalence in any single conformation, and PAM derived by

Table 3. Molecular Polarizability (\AA^3) of Aromatic Systems^a

		α_x	α_y	α_z
Benzene	Expt	11.70	11.70	5.72
	 Expt	12.26	12.26	6.66
	AMOEBA	12.30	12.30	6.64
Naphthalene	Expt	20.20	18.80	10.70
	Expt	22.20	18.20	7.30
	AMOEBA	21.78	18.51	9.77
Anthracene	Expt	35.20	25.60	15.20
	 Expt	44.70	25.80	9.80
	DFT(B3LYP/6-31G*)	38.65	21.65	6.51
	AMOEBA	32.85	24.67	12.63
Nanotube Armchair (3,3)	DFT(B3LYP/6-31G*)	59.65	39.06	39.06
	 DFT(B3LYP/6-31+G*)	64.45	47.33	47.33
	AMOEBA	61.68	41.20	41.20

^a Experimental data are taken from Table 8 of Applequist.¹⁴⁸

simple averaging would lead to a large error in the molecular dipole moment. The potential-optimized PAM, where methyl hydrogens are constrained to adopt equivalent values, will reproduce almost exactly both the *ab initio* potential and the molecular multipole moments. Our standard procedure is to use a molecular potential grid consisting of a 2 \AA shell beginning 1 \AA out from the vdW surface. The DMA monopole values are generally fixed during the potential fitting procedure.

This electrostatic parametrization protocol is particularly important for larger molecules and for molecules with high symmetry. It is known that the original DMA approach tends to give “unphysical” multipole values for large molecules when diffuse functions are included in the basis set even though the resulting electrostatic potential is correct. A recent modification of DMA⁸⁰ has been put forward to address this issue. However, in our hands, the multipoles from the modified scheme seem less transferable between conformations. The above protocol allows derivation of PAM corresponding to larger basis sets than would be practical with the original DMA method. Note that this procedure is different from restrained potential fits commonly used to fit fixed atomic charge models, as the starting DMA values are already quite reasonable and the fitting can be considered as a small perturbation biased toward the larger basis set potential. The overall procedure has been extensively tested in a small molecule hydration study⁸¹ and will be used in future AMOEBA parameterization efforts.

Empirical vdW parameters were determined by fitting to both gas and liquid phase properties. The gas phase properties include homodimer binding energy (BSSE corrected) and structure from *ab initio* calculations at the MP2/aug-cc-pVTZ level or above. Liquid properties include experimental density and the heat of vaporization of neat liquids. The vdW parameters were first estimated by comparing the structure and energy of the AMOEBA-optimized dimer with *ab initio* results and then fine-tuned to

reproduce the experimental liquid density and heat of vaporization via molecular dynamics simulation. Additional homodimers at alternative configurations, heterodimers with water, and liquid properties were computed *post facto* for the purpose of validation. A more generic force field atom classification for vdW parameters was enforced to ensure transferability. Table 1 lists the common vdW atom classes used by AMOEBA, together with the corresponding vdW parameters and polarizabilities. The vdW atom classes are also used to define parameters for all of the valence potential energy terms. The parameters for bonded terms, initially transferred from MM3, are optimized to reproduce *ab initio* geometries and vibration frequencies. In the final parametrization step, after all other parameters are fixed, torsional parameters are obtained by fitting to *ab initio* conformational energy profiles at the MP2/6-311++(2d,2p) level of theory.

COMPUTATION DETAILS

All force field calculations were carried out using the TINKER molecular modeling package.⁸² The *ab initio* molecular orbital calculations were performed using Gaussian 03.⁷⁸ AMOEBA energy minimization of gas phase dimers was performed to achieve a RMS gradient of 0.01 kcal/mol/Å per atom. For bulk phase simulations, Particle Mesh Ewald was applied to treat the long-range electrostatic interactions, with a 9 Å real-space cutoff. A 12 Å switched cutoff is used for vdW interactions. Crystal minimizations were terminated when the gradient fell below 0.1 kcal/mol/Å per degree of freedom (atomic coordinates and cell parameters). The periodic box for neat liquid simulation systems was a cube approximately 20 Å on a side. A 2 ns NVT simulation was performed for each neat liquid, with an integration time step of 1 fs and the density set to the experimental value. A Berendsen thermostat was used to control the temperature.⁸³ The average pressure, heat of vaporization, and diffusion constant were calculated from the trajectories using the same formula reported for water.¹⁸ Starting from the final configuration of each NVT run, NPT simulations of 2 ns were performed and used to compute average densities. NPT simulations of up to 6 ns were used to estimate the dielectric constants for selected compounds. NPT simulations used the Berendsen barostat with a relaxation time τ of 5 ps to control pressure.⁸³ Solvation free energies were computed using the same free energy perturbation procedure reported previously.⁸⁴ The hydration free energy of each small molecule was calculated by summing up free energies for three thermodynamic cycle steps: solute discharging in a vacuum, solute vdW coupling with a solvent (water), and solute recharging in water. Charging steps were performed over seven windows, and vdW coupling steps were performed in 16 windows. A softcore modification of the buffered-14-7 function was used in the vdW coupling.⁸⁴ Samples of solutes in a vacuum were collected every 0.5 ps from 10 ns stochastic dynamics simulations with an integration time step of 0.1 fs. Condensed phase simulations were run for 1 ns under NVT in 850 water molecules as the solvent, with the system density fixed at 1.000 g cm⁻³. Snapshots were saved every 0.5 ps for free energy evaluation. Induced dipoles are converged to a RMS change of 0.00001 D per step for simulations in a vacuum and 0.01 D in bulk simulations. Free energy was calculated by re-evaluating the energy from the saved MD snapshots with the induced dipoles converged to 0.00001 D RMS. The Bennett Acceptance Ratio (BAR) method⁸⁵ was used to estimate the free energy change between neighboring steps.

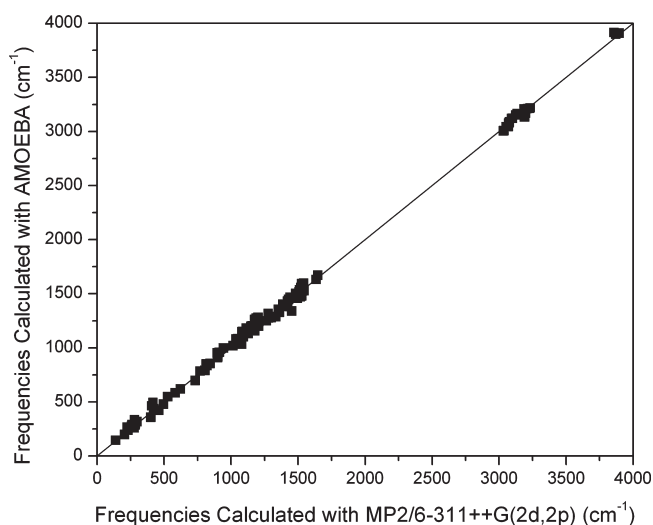


Figure 2. The comparison of gas-phase vibration frequencies calculated via MP2/6-311++G(2d,2p) and the AMOEBA force field. The signed average error is -9.59 cm^{-1} ; the unsigned average error is 27.17 cm^{-1} . The RMSE is 33.77 cm^{-1} . Molecules included are methanol, ethanol, propanol, dimethylether, and phenol.

The TINKER VIBRATE program was used for normal mode calculations, implemented via diagonalization of the mass-weighted Hessian matrix of Cartesian second derivatives.

RESULTS AND DISCUSSION

Gas Phase Calculations. In conventional fixed charge force fields, atomic partial charges are often “pre-polarized” to match those in the liquid state by empirical scaling of *ab initio* charges from the gas phase. In contrast, the electrostatics parameters in AMOEBA are derived from high-level *ab initio* calculations in the gas phase, and electronic polarization by the environment is accounted for explicitly. Homodimer association energies and equilibrium geometries in the gas phase were used in conjunction with the condensed-phase properties to obtain vdW parameters for alkanes, aromatics, amines, alcohols, amides, and sulfides. Additional local minima corresponding to heterodimers with a water molecule, as hydrogen bond donor and acceptor, were utilized to further validate the parameters.

Intramolecular Interactions. The intramolecular valence parameters for AMOEBA were initially transferred from MM3.¹ These values were already known to perform satisfactorily in terms of producing reasonable equilibrium molecular geometries. We have further optimized the bond, angle, and other valence parameters against the *ab initio* equilibrium structures and vibration frequencies. As an illustration, the vibration frequencies of methanol, ethanol, propanol, dimethylether, and phenol (total of 117 data points) calculated using the final AMOEBA parameters are compared with MP2/6-311++G(2d,2p) data in Figure 2, where excellent agreement between AMOEBA and QM is seen to result from the parameter optimization. The average absolute vibration frequency error is 27.17 cm^{-1} , with a root-mean-square error (RMSE) of 33.77 cm^{-1} . The inclusion of a stretch–bend cross term is critical to the quality of the vibration frequencies. Examples of conformational energy are given in Figure 3 for butane, methanol, phenol, ethylamine, and ethylsulfide. The AMOEBA intramolecular nonbonded interactions, plus a standard torsional energy contribution computed via a

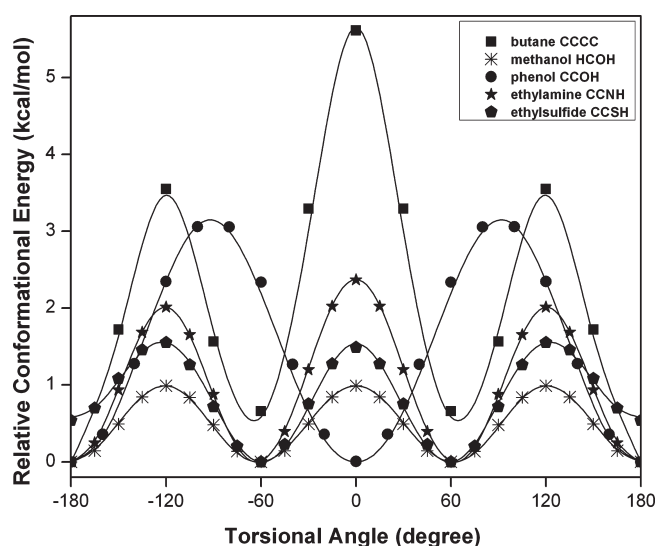


Figure 3. Relative conformational energies with respect to specific torsions. Solid line, MP2/6-311++G(2d,2p); symbols, AMOEBA. RMSEs between AMOEBA and QM results: butane CCCC, 0.14 kcal/mol; methanol HCOH, 0.016 kcal/mol; phenol CCOH, 0.069 kcal/mol; ethylamine CCNH, 0.014 kcal/mol; ethylsulfide CCSH, 0.031 kcal/mol.

Table 4. Relative Conformational Energies (kcal/mol) of *n*-Butane

	AMOEBA	<i>ab initio</i> ^a	experimental ^b
anti	0.00	0.00	0.00
syn	5.61	5.50	3.95
gauche	0.51	0.62	0.67
120°	3.55	3.31	3.62

^a Ref 149. ^b Ref 150.

traditional three-term Fourier expansion, reproduce the *ab initio* conformational energy very well for these and other organic compounds. Table 4 gives the relative conformational energies (kcal/mol) of *n*-butane.

Intermolecular Interactions. The equilibrium geometry and association energy of 32 homo- and heterodimers of alkanes, amides, alcohols, amines, sulfides, and aromatics have been computed. Results are reported in Table 5. For homodimers, the global minima have been utilized in the parametrization process, with vdW parameters adjusted further on the basis of liquid properties (see below). The results reported here are computed using the final AMOEBA parameter values. In addition, heterodimers with a water molecule serve as a test of the transferability of the model. The parameters of the water model were as reported previously.¹⁸ In Table 5, the dimer equilibrium geometries and association energies from AMOEBA and MP2/aug-cc-pVTZ, with and without basis set superimposition error (BSSE) correction, are compared. The excellent correlation ($R^2 = 0.99$) of the AMOEBA energy with BSSE corrected *ab initio* QM association energy is shown in Figure 4. Across the 32 dimer pairs, the average unsigned error in AMOEBA association energy is 0.31 kcal/mol with a RMSE of 0.38 kcal/mol. The AMOEBA association energy is calculated from the AMOEBA-optimized dimer structures, which are also in good agreement with *ab initio* structures, as indicated by the hydrogen bond distance and angle

comparison in Table 5. In comparison to the aug-cc-pVTZ dimer geometry, AMOEBA gives a 0.056 Å RMSE in hydrogen bond lengths and 9.53° in angle values. Recently, Faver et al.⁸⁶ compared AMOEBA results for a series of dimer interaction energies. They found AMOEBA energies to be in better agreement with high-level QM results than the GAFF and MMFF force fields as well as many lower-level QM protocols. The CHARMM fluctuating charge force field, another systematically derived polarizable force field, reported a 0.19 Å RMSE in hydrogen bond distance and 0.98 kcal/mol in dimerization energy for a series of solute–water complexes when compared to DFT results.²⁹ Harder et al. reported good agreement on NMA–water dimer energies with MP2/6-311+G(3df,2p) using a Drude oscillator base polarizable force field.³⁴

In addition to the global minimum structures discussed above, several additional local minima for the dimers of formamide, DMF, NMA–water, ammonia, and benzene have been investigated using AMOEBA. These calculations provide an additional check of the potential energy surface beyond the global energy basin.

Hydrogen Bond Directionality. One of the most important interactions in organic molecules, and one that molecular mechanics methods should model accurately, is the hydrogen bond. Classical molecular orbital arguments describe the directional dependence of hydrogen bonding as a balance between electrostatics and charge transfer.⁸⁷ Since the mid-1980s,⁸⁸ most potentials for biological simulation have used simple Coulombic interactions to describe hydrogen bonds, while some organic force fields have incorporated special directionally dependent terms.⁸⁹ Recently, a new energy decomposition analysis of the water dimer based on absolutely localized molecular orbitals (ALMOs) by the Head-Gordon group suggested that the water–water interaction is largely due to electrostatics and polarization, with only minimal formal charge transfer.⁹⁰ In Figure 5, we compare AMOEBA results for the directionality of the formaldehyde–water hydrogen bond with those from MP2/aug-cc-pVTZ calculations and the fixed partial atomic charge-based OPLS-AA force field. For the planar hydrogen bonded structures, both AMOEBA and *ab initio* calculations yield minima at an acceptor angle near 120° with the linear structure lying about 1.5 kcal/mol higher in energy. This is in rough agreement with statistical distributions compiled from small molecule X-ray structures.⁹¹ For OPLS-AA and other fixed partial charge models such as Amber and CHARMM (data not shown), there is very little angular dependence of the energy for a broad range of values centered at 180°. In AMOEBA, the quadrupole value on the carbonyl oxygen plays a major role in favoring the nonlinear configuration. A recently reported NEMO potential for formaldehyde provides independent evidence for the importance of local quadrupole moments.⁹² Models based entirely on atomic charges (or atomic dipoles) are unable to break the symmetry along the C=O axis in order to provide a more favorable electrostatic potential at the “lone pair” angles. While the directional dependence of atomic charge models can be improved by including additional charges at lone pair⁹³ or π -cloud sites,⁹⁴ this is not a general solution and may not adequately address nonstandard hydrogen bonds.^{95,96}

Amides. *Ab initio* studies of a series of formamide and dimethylformamide dimers have been reported previously^{97,98} and are compared with our own quantum calculations and AMOEBA results in Table 6. In earlier work, we reported the inter- and intramolecular electrostatic models for NMA and alanine dipeptide,¹⁵

Table 5. Gas Phase Dimer Equilibrium Structure and Binding Energy from QM and AMOEBA

dimer	bond dist (Å) ^a /angle (degree) ^b		binding energy (kcal/mol)		
	MP2/aug-cc-pVDZ	AMOEBA	MP2/aug-cc-pVTZ ^c	BSSE corrected	AMOEBA
methane–water	3.49/86.90	3.48/84.90	−1.18	−0.92	−1.21
methane–methane	4.01/179.96	3.93/165.00	−0.52	−0.36	−0.53
methanol–WD ^d	2.84/165.73	2.83/174.14	−6.10	−5.51	−5.85
methanol–WA ^e	2.90/177.05	2.99/178.30	−5.30	−4.78	−4.77
methanol–methanol	2.85/167.78,179.80 ^f	2.88/179.42	−6.33	−5.26	−5.66
Ethanol–WD	2.84/161.12	2.87/175.48	−6.32	−5.70	−5.65
ethanol–WA	2.91/177.12	2.93/179.81	−5.27	−4.71	−4.67
ethanol–ethanol	2.86/166.46	2.89/167.63	−6.43	−5.62	−5.80
isopropanol–WD	2.84/161.68	2.89/168.37	−6.85	−6.11	−5.87
isopropanol–WA	2.92/177.72	2.89/176.15	−5.47	−4.85	−5.28
dimethylether–WD	2.81/159.96	2.84/174.51	−6.68	−5.93	−6.28
phenol–WD	3.01/163.55	2.94/167.73	−4.78	−3.98	−4.58
phenol–WA	2.84/176.64	2.88/176.95	−7.45	−6.71	−6.32
<i>p</i> -cresol–WD	3.00/162.90	2.92/167.62	−4.91	−4.14	−4.79
<i>p</i> -cresol–WA	2.85/176.76	2.88/176.85	−7.28	−6.55	−6.46
H ₂ S–WD	3.49/164.81	3.36/170.65	−3.35	−2.89	−3.48
H ₂ S–WA	3.54/176.38	3.60/173.92	−3.04	−2.70	−2.78
H ₂ S–H ₂ S	4.09/172.62	4.04/167.90	−2.18	−1.81	−2.09
methylsulfide–WD	3.33/150.58	3.28/164.20	−4.94	−4.34	−4.84
methylsulfide–WA	3.57/170.47	3.62/176.74	−2.73	−2.38	−2.52
dimethylsulfide–WD	3.25/150.17	3.23/166.87	−6.07	−5.36	−5.19
methylamine–WD	2.86/162.25	2.86/175.87	−8.09	−7.45	−8.46
methylamine–methylamine	3.16/153.85	3.20/157.57	−4.74	−4.14	−4.09
ethylamine–WD	2.87/162.09	2.91/173.27	−8.17	−7.49	−7.57
imidazole–WA	2.87/160.21	2.93/177.67	−8.34	−7.05	−7.68
indole–WA	2.96/179.99	3.04/171.55	−6.52	−5.78	−5.58
ethylsulfide–WD	3.32/151.03	3.23/165.79	−5.52	−4.84	−5.46
ethylsulfide–WA	3.57/168.35	3.66/172.76	−2.69	−2.33	−2.10
methylethylsulfide–WD	3.24/151.66	3.20/168.66	−6.46	−5.68	−5.78
formamide–WD	1.91/99.52	1.86/115.00	−7.32	−6.75	−6.94
formamide–formamide	1.84/174.28	1.87/176.92	−16.86	−15.62	−16.00
NMA–WD	1.85/165.32	1.82/172.52	−8.71	−7.98	−8.38

^a Heavy atom distance in the hydrogen bond, link O–O or O–N. ^b Hydrogen bond angle N(O)–H···N(O) except for methane. ^c Single point with MP2/aug-cc-pVTZ after structural optimization with MP2/aug-cc-pVDZ. ^d WD denotes water as the hydrogen bond donor in dimer structure. ^e WA denotes water as the hydrogen bond acceptor in the dimer structure. ^f MP2/6-31+G* optimization result.

as well as free energy calculations of ion solvation in liquid formamide.⁹⁹

The six formamide dimer configurations investigated here are similar to those found by Vargas et al.⁹⁷ The global minimum is a cyclic configuration where two O···H hydrogen bonds constrain the dimer to form an 8-membered ring. The hydrogen bonding structure and distance in this dimer were used to adjust vdW parameters for carbonyl oxygen and amide hydrogen, while liquid simulations of a series of amides were utilized to fine-tune vdW parameters for the amide C, O, N, and H atoms simultaneously. As shown in Table 6, the six formamide dimer configurations optimized using MP2/aug-cc-pVDZ and AMOEBA are in excellent agreement, with an average RMSE of ~0.1 Å in atomic coordinates.

For comparison, association energies were computed using AMOEBA and at the MP2/aug-cc-pVQZ level (for MP2/aug-cc-pVDZ optimized geometries) for each configuration. At this level of *ab initio* theory, the BSSE correction is about 0.6 kcal/mol for

dimer A. Sponer and Hobza⁹⁸ have shown that the association energy is well converged between the MP2/aug-cc-pVQZ and MP2/cc-pVSZ levels. At the MP2/aug-cc-pVTZ level, the association energy of dimer A is about 0.5 kcal/mol less negative than that obtained with the aug-cc-pVQZ basis set. Agreement between the *ab initio* and AMOEBA association energies for all six dimer configurations is 0.32 ± 0.22 kcal/mol.

These same dimers, especially the cyclic structure A, have been widely investigated using various *ab initio* methods, including CCSD(T)/CBS.^{98,100,101} Both association and binding energies have been reported. The *association* energy is defined as the energy to separate the dimer without relaxing the monomer geometry, while the *binding* energy refers to the energy of the dimer relative to those of fully relaxed monomers. The difference between the two, i.e., the deformation energy upon binding, is reasonably small except for the cyclic dimer A. Our calculations suggest a deformation energy per molecule in dimer A of 0.8 kcal/mol at the MP2/aug-cc-pVQZ level. Moving from the

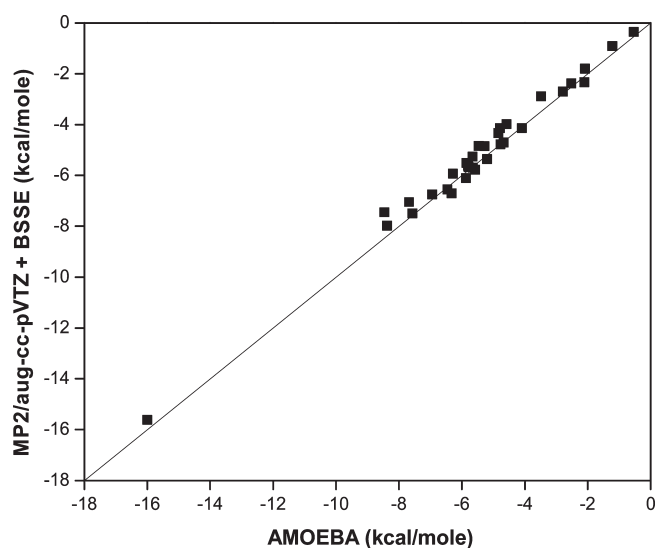


Figure 4. Comparison of dimer binding energies given by BSSE-corrected QM results and AMOEBA calculations. A total of 32 dimer structures are included, with a signed error of 0.22 kcal/mol, unsigned error of 0.31 kcal/mol, and RMSE of 0.38 kcal/mol.

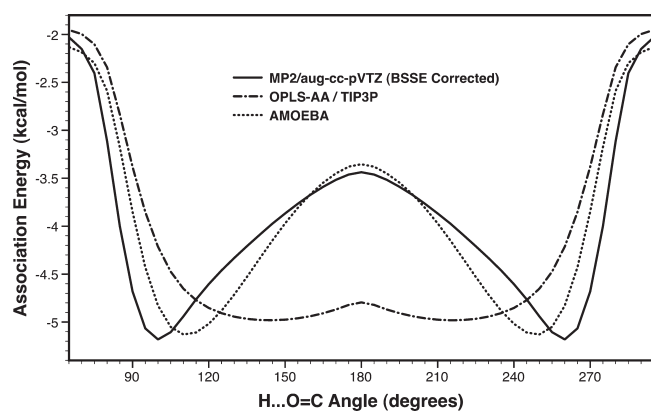


Figure 5. Association energy for the hydrogen bonded formaldehyde–water dimer as a function of the $\text{H}\cdots\text{O}=\text{C}$ angle. OPLS-AA/TIP3P is an OPLS-AA model for formaldehyde with a TIP3P water molecule. All energies are for structures fully optimized with a constrained $\text{H}\cdots\text{O}=\text{C}$ angle and with all atoms lying in a plane. The curves shown are interpolated from discrete calculations performed with each of the three methods at angle intervals of $5\text{--}15^\circ$.

equilibrium geometry of monomeric formamide to that in dimer A, the $\text{C}=\text{O}$ bond elongates and the amide $\text{C}\text{--}\text{N}$ bond becomes shorter, similar to the changes observed for amides upon moving from the gas to the liquid phase. The strong orbital delocalization favored by dimer A leads to changes in bond order, and the large deformation energy observed is not correctly described by the classical description of bond stretching and angle bending used by AMOEBA. In the force field calculations, the changes in the bond lengths are in the correct direction; however, the magnitudes are less than those observed in *ab initio* results. Thus, the AMOEBA deformation energy for dimer A is only 0.2 kcal/mol per molecule. Similar problems exist in the treatment of general conjugated, π -bonded systems. One possible solution, as implemented for the MM3 model,¹⁰² involves using a simple VESCF

Table 6. Gas Phase Dimer Association Energy (kcal/mol) and Structure (Å) from *ab Initio* and AMOEBA Calculations of Multiple Configurations

	<i>ab initio</i> QM	AMOEBA	struct. RMSE ^d	
formamide ^b				
A (cyc)	−16.1, −16.1, ^c −15.96 ^d	−16.0	0.03	
B (s1)	−10.6	−10.3	0.05	
C (np3)	−8.2	−8.9	0.07	
D (np1)	−7.2	−7.5	0.23	
E (s2)	−6.9	−7.3	0.04	
F (HT)	−5.4	−5.5	0.08	
DMF ^e				
	w/o BSSE	w/BSSE		
A	−6.95	−5.35	−5.01	0.08
B	−5.82	−4.14	−5.62	0.08
C	−11.41	−8.34	−7.39	0.26
D	−12.11	−8.90	−8.25	0.15
NMA–water ^f				
A	−8.07	−8.34	0.14	
B	−8.01	−8.13	0.05	
C	−5.18	−5.22	0.13	
NH_3 ^b				
linear	3.03	3.20	0.14	
asymmetrical	3.07	3.21	0.17	
benzene ^g				
T	−2.57 ^h	−2.74 ⁱ	−2.16	0.10
TT	−2.66	NA	−2.61	0.15
PD	−2.49	−2.78	−2.80	0.04
S	−1.51	−1.81	−2.05	0.13

^aThis study. RMS deviations between the AMOEBA and MP2/aug-cc-pVDZ optimized dimer structures; methyl hydrogen atoms excluded. ^bData from this study. MP2/aug-cc-pVQZ with BSSE correction. ^cRef 97. Geometry optimized at the MP2/aug-cc-pVTZ level. The binding energy was reported, which was adjusted to an association energy based on a deformation energy of 1.6 kcal/mol total. ^dRef 101. CCSD(T)/CBS results for association energy. ^eRef 97. MP2/aug-cc-pVTZ. ^fData from this study. MP2/aug-cc-pVTZ with BSSE correction. ^gBenzene dimer structures: T-shaped (T), T-shaped tilted (TT), parallel displaced (PD), and parallel sandwich (S). ^hRef 109. DFT-D structures and association energy from CCSD(T)₂[T 70%] results. ⁱRef 108. Estimated CCSD(T)/CBS.

molecular orbital calculation to reassess bond orders on the fly. In the current work, we omit such MO-based corrections and attempt to compromise between agreement with the gas phase dissociation energies and reproduction of liquid phase thermodynamics, structure, and dynamics.

The dimethylformamide dimer provides an additional validation of the amide vdW parameters. *Ab initio* results, including binding energy and dimer structures at equilibrium, were reported by Vargas et al.⁹⁵ A comparison between AMOEBA and the *ab initio* results is made in Table 6. It should be noted that the BSSE corrections at the MP2/aug-cc-pVTZ level are almost 2 kcal/mol for dimers A and B, and more than 3 kcal/mol for dimers C and D. Overall, the AMOEBA results closely follow the BSSE corrected *ab initio* values.

Furthermore, three configurations of *N*-methylacetamide (NMA) complexed with a single water molecule have been identified as minima by MP2/6-31+G* energy minimization, similar to those

Table 7. Formic Acid Gas Phase Dimer Energy (kcal/mol) and Structure (Å) from *ab Initio* and AMOEBA Calculations

	<i>ab initio</i> QM ^a		AMOEBA	struct
	E_{assoc}	E_{bind}	E_{assoc}	RMSE
A	-18.6	-15.8	15.9	0.05
B	-10.3	-9.6	-9.7	0.04
C	-6.7	-6.1	-6.5	0.03
D	-3.4	-3.2	-3.5	0.06
E	-2.4	-2.2	-2.5	0.03
F	-4.4	-4.2	-4.4	0.05

^a MP2/aug-cc-pVQZ energy at the MP2/aug-cc-pVDZ optimized geometry. Configurations A–F are analogous to those in Table 6 for the formamide dimer.

reported previously.²⁹ The association energies were evaluated at the MP2/aug-cc-pVTZ level. In two of the minima, water is the hydrogen donor, while in the third the water oxygen atom is hydrogen bonded to the NMA amide hydrogen. In Table 6, the association energy and equilibrium structure given by AMOEBA are shown to be in excellent agreement with *ab initio* results for all three configurations.

Amines, Alcohols, and Acids. Among the stationary point structures of the ammonia dimer explored in the literature,^{103,104} the so-called “linear” configuration is found to be a true minimum by both AMOEBA and *ab initio* optimization. A comparison of energies and geometries is given in Table 6. Analogous to AMOEBA water,¹⁸ the quadrupole components of the ammonia N and H atoms are scaled downward by 40%. The scaling of quadrupole moments has no effect on the molecular dipole moment but is necessary for producing correct dimer equilibrium structures (i.e., “flap angles”) in the case of both water and ammonia, possibly due to the limited basis set used in the PAM derivation or the lack of quadrupole polarization in the AMOEBA model. For ammonia, we have chosen a quadrupole scaling factor of 0.6, which yields a reduced molecular quadrupole moment ($Q_{zz} = 2.46$ au) in close agreement with available experimental values of 2.42 ± 0.04 au¹⁰⁵ and 2.45 ± 0.30 au.^{106,107} For consistency, we have similarly scaled the atomic quadrupole moments of the –NH and –OH groups in amines and alcohols so they are comparable to those of water and ammonia. AMOEBA-predicted methylamine and methanol dimer structures and energies are compared to corresponding *ab initio* results in Table 6. Results from liquid simulations are also satisfactory, as shown in the next section on condensed-phase simulations.

For both aldehydes and acids, the vdW parameters of the carbonyl group are transferred from amides, as given in Table 1. A number of dimer configurations for formic acid have been investigated using both AMOEBA and *ab initio* methods. In Table 7, we note that the deformation energy upon binding is 1.4 kcal/mol for formic acid configuration A (even higher than that of formamide), while in other configurations it is only a fraction of a kilocalorie per mole. This suggests a very strong electron delocalization in the cyclic configuration of dimer A. The BSSE from *ab initio* binding calculation using MP2/aug-cc-pVQZ at the MP2/aug-cc-pVTZ minimum geometry is 0.8 kcal/mol for dimer A. Deformation energies under the AMOEBA model are close to 0.1 kcal/mol, as expected.

Benzene and Other Aromatics. Benzene dimers have been widely studied in recent years using very high-level *ab initio* quantum mechanics.^{108–111} A comparison between AMOEBA

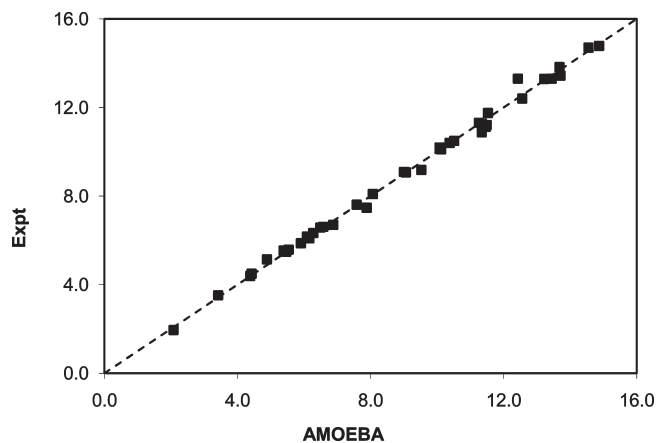


Figure 6. Comparison of heat of vaporizations (kcal/mol) from experimental results and from liquid simulations with AMOEBA.

and QM results on four dominant stationary benzene dimer configurations is shown in Table 6. Overall agreement for the structure and energy is satisfactory, although AMOEBA selects the PD (parallel-displaced) structure as the global minimum while QM favors the TT (T-shaped tilted) configuration. As has been noted, the potential energy surface of the benzene dimer is extremely shallow.¹⁰⁹ According to AMOEBA, the two T-shaped dimers (T and TT) are transition state structures rather than true minima, as indicated by negative eigenvalues of the Hessian matrix. Many QM calculations impose symmetry and/or do not minimize completely, so it is not yet clear which structures are true minima on high-level *ab initio* surfaces.

The polarizability of various aromatics and a carbon nanotube section were calculated using AMOEBA with an atomic polarizability of 1.75 \AA^3 for C and 0.696 \AA^3 for H. The carbon nanotube studied is an “armchair” configuration made of three units of (3,3) structure. DFT calculations of molecular polarizability were performed using geometries optimized at the HF/6-31G* level. Many combinations of carbon and hydrogen atomic polarizabilities are able to reproduce the benzene molecular polarizability quite reasonably. However, upon comparing the molecular polarizabilities of a series of aromatic compounds, it is found that aromatic atomic polarizabilities for carbon and hydrogen must be increased from aliphatic values (see Table 3).

Condensed Phase Simulations. *Density and Heat of Vaporization.* The experimental density and heat of vaporization of a series of neat organic liquids were used to optimize vdW parameters. To enforce transferability, sets of compounds sharing the same atom types (Table 1) were parametrized together with compounds from different functional group families. This simultaneous parametrization helps to maintain the chemical consistency among different elements and functional groups. Liquid phase MD simulations were used to sample virial-based pressure values from NVT ensembles, and at the same time the heat of vaporization was computed from these trajectories. For liquids with low compressibility, such as water (5.1×10^{-5} per bar at 273 K), a pressure of 200 atm corresponds roughly to a 1% change in density. Thus, a reasonable target for molecules under AMOEBA was taken as an average pressure within the range 1 ± 200 atm, while keeping the heat of vaporization within ± 0.5 kcal/mol of the experimental result. See Figure 6 for a comparison of heat of vaporizations (kcal/mol) from experimental results and from liquid simulations with AMOEBA.

Table 8. Heat of Vaporization (kcal/mol) and Pressure from NVT Simulations of Neat Liquids^a

	E_{liq}	E_{gas}	ΔH_{sim}	ΔH_{exptl}	P_{sim}	T	ρ_{exptl}
water	-9.02	0.90	10.51	10.49 ^b	-61	298.2	0.997 ^b
MeOH	-3.57	4.90	9.06	8.95 ^c	-40	298.2	0.786 ^c
EtOH	-3.73	5.80	10.12	10.11 ^c	-40	298.2	0.785 ^c
<i>n</i> -PrOH	-1.41	9.26	11.26	11.31 ^c	-42	298.2	0.800 ^c
<i>i</i> -PrOH	-2.68	8.07	11.34	10.88 ^c	-21	298.2	0.781 ^c
NH ₃	-2.89	2.17	5.54	5.58 ^e	146	239.8	0.682 ^e
MeNH ₂	-2.89	2.67	6.09	6.17 ^f	177	266.9	0.694 ^g
EtNH ₂	1.94	8.24	6.88	6.70 ^h	-46	289.7	0.687 ⁱ
PrNH ₂	4.24	11.53	7.89	7.47 ^b	-184	298.2	0.711 ^j
di-Me amine	3.73	9.45	6.28	6.33 ^k	-97	280.0	0.671 ⁱ
tri-Me amine	15.88	20.95	5.62	5.48 ^l	-231	276.0	0.653 ⁱ
formic acid	-16.04	-5.17	11.46	11.13 ^m	163	298.2	1.214 ^c
acetic acid	-22.88	-10.91	12.56	12.49 ^d	8	298.2	1.044 ⁿ
formaldehyde	-3.47	1.41	5.38	5.54 ^{b,o}	-65	254.0	0.812 ^{b,o}
acetaldehyde	-6.73	-1.15	6.16	6.09 ^{b,o}	184	293.2	0.783 ^{b,o}
di-Me ether	3.00	7.39	4.89	5.14 ^{b,o}	-62	248.3	0.736 ^{b,o}
H ₂ S	-3.28	0.67	4.38	4.39 ^p	-12	220.2	0.934 ^p
MeSH	-2.95	2.40	5.91	5.87 ^q	180	280.0	0.891 ^r
EtSH	2.00	7.90	6.49	6.58 ^d	109	298.1	0.833 ^s
di-MeS	0.15	6.14	6.58	6.61 ^c	28	298.2	0.842 ^t
di-MeS ₂	-7.39	1.54	9.53	9.18 ^d	-15	298.2	1.057 ^s
MeEtS	0.98	7.97	7.58	7.61 ^c	20	298.2	0.837 ^s
benzene	10.90	18.38	8.07	8.09 ^{u,v}	96	298.0	0.874 ^{u,v}
toluene	3.79	12.20	9.01	9.09 ^b	142	298.2	0.865 ^b
phenol	-5.87	7.22	13.68	13.82 ^c	-52	298.2	1.058 ^c
phenol	-4.44	8.14	13.22	13.36 ^{b,o}	270	323.0	1.050 ^{b,o}
ethylbenzene	10.68	20.17	10.08	10.10 ^b	93	294.0	0.863 ^b
cresol	-2.98	11.27	14.87	14.77 ^w	173	313.2	1.019 ^b
formamide	-17.96	-4.00	14.55	14.70 ^x	-18	298.2	1.129 ^y
<i>N</i> -MeForm	-12.57	0.54	13.71	13.43 ^o	52	298.2	1.005 ^o
di-MeForm	-5.86	5.04	11.49	11.21 ^{b,o}	-129	298.2	0.944 ^{b,o}
di-MeForm	-2.36	7.28	10.38	10.40 ^z	107	373.2	0.873 ^{aa,bb,cc}
acetamide	-18.88	-4.92	14.70	14.23 ^w	6	373.0	0.984 ^w
acetamide	-13.87	-2.42	12.43	13.30 ^w	198	494.2	0.867 ^w
<i>N</i> -MeAcet	-13.33	-0.62	13.44	13.30 ^{dd}	30	373.2	0.894 ^{dd}
di-MeAcet	-8.78	2.16	11.53	11.75 ^c	349	298.2	0.936 ^c
methane	-0.86	1.00	2.08	1.95 ^{u,v}	293	111.0	0.424 ^{u,v}
ethane	1.81	4.86	3.42	3.52 ^{u,v}	98	184.0	0.546 ^{u,v}
propane	5.24	9.21	4.43	4.49 ^{u,v}	19	230.0	0.581 ^{u,v}

^a E is the potential energy in kcal/mol. ΔH is the heat of vaporization in kcal/mol, calculated as $\Delta H = E_{\text{gas}} - E_{\text{liq}} + RT$. P is pressure in atmospheres. T is temperature in Kelvin. ρ is density in g cm^{-3} . ^b Ref 151. ^c Ref 152. ^d Ref 153. ^e Ref 154. ^f Ref 155. ^g Ref 156. ^h Ref 157. ⁱ Ref 158. ^j Ref 159. ^k Ref 160. ^l Ref 161. ^m Ref 162. ⁿ Ref 163. ^o Ref 164. ^p Ref 165. ^q Ref 166. ^r Ref 167. ^s Ref 168. ^t Ref 169. ^u Ref 170. ^v Ref 171. ^w Ref 172. ^x Ref 173. ^y Ref 174. ^z Ref 175. ^{aa} Ref 176. ^{bb} Ref 177. ^{cc} Ref 178. ^{dd} Ref 179.

In Table 8, the results from liquid simulations are compared with experimental results. The overall RMSE in the heat of vaporization is 0.23 kcal/mol for the 37 compounds listed. The largest error of 0.9 kcal/mol was observed for acetamide at 494 K. The average pressure from the NVT simulations of all 37 compounds is 39 ± 124 atm, and the RMSE from the experimental

Table 9. Comparison of Hydrogen Sulfide Liquid Properties from Experimental Results¹⁶⁵ and AMOEBA Simulation Results

T (K)	E_{liq}	E_{gas}	ΔH_{expt}	ΔH_{sim}	P_{expt}	P_{sim}	ρ_{expt}
220.2	-3.28	0.67	4.39	4.38	1.441	-12	0.934
239.7	-3.00	0.73	4.19	4.20	3.326	54	0.902
252.4	-2.82	0.76	4.03	4.08	5.321	83	0.878
281.2	-2.39	0.85	3.63	3.80	12.969	163	0.816

value (1 atm) is 131 atm. This pressure deviation corresponds to a less than 1% change in density for most of these liquids. Selected NPT simulations were performed to confirm the density estimates. For liquid ammonia, we obtained an average pressure of 146 atm from the NVT simulation at the experimental density 0.682 g cm^{-3} , while a corresponding NPT trajectory at 1 atm gave an average density of 0.676 g cm^{-3} (relative error 0.8%). For formic acid, the liquid expanded slightly from 1.218 to 1.200 g cm^{-3} when the pressure changed from 163 to 1 atm (1.5%). For methanol, the density increased by 0.3% from the 0.786 g cm^{-3} experimental value in the NPT simulation, while the NVT simulation produced an average pressure of -40 atm. The largest error is for the NVT pressure of 349 atm for dimethylacetamide (DMA), resulting in a density decrease of 2.2% from 0.936 to 0.915 in the NPT simulation. The statistical error in the pressure, estimated using a block average approach, is on the order of 50 atm, while the statistical error in the heat of vaporization is negligible. For liquid hydrogen sulfide, the heat of vaporization values calculated at four different temperatures are in excellent agreement with experimental values, as shown in Table 9.

The overall performance on neat liquid properties seems slightly better than for fixed charge potentials such as OPLS-AA⁴ and COMPASS.^{112,113} Density and heat of vaporization are explicit targets in the AMOEBA force field optimization and development, as they are in nearly all force field models intended for bulk simulation. However, it should be kept in mind that gas-phase cluster properties computed using the same AMOEBA parameter set are also in good agreement with *ab initio* MP2 results. The error in heat of vaporization for 14 organic liquids given by the CHARMM fluctuating charge force field was about 1 kcal/mol, which is somewhat greater than the fixed charge CHARMM force field.²⁹ Other polarizable force fields, including Drude-oscillator^{34,35} and PIPF models,²² have reported an accuracy comparable to AMOEBA for selected molecules.

Dielectric and Diffusion Constants. Dielectric constants and diffusion constants were computed for selected compounds and are compared with available experimental measurements in Figures 7 and 8 (see also Table 10). For both static dielectric and self-diffusion constants, the overall agreement between AMOEBA and experiment values is satisfactory. Static dielectric constants were computed numerically from the cell dipole moment fluctuations and generally required multiple nanoseconds of simulation time to achieve convergence. Dielectric constants of individual liquids have been reported previously from fixed charge and polarizable force field simulations. Among the common fixed-charge water models, TIP5P⁵⁴ reproduces the experimental static dielectric constant accurately, while other TIPxP,^{53,114} SPC,¹¹⁵ and SPC/E¹¹⁶ models give values that are either far too low or much too high.^{54,117} There seems to be no obvious correlation between the molecular electric moments and the

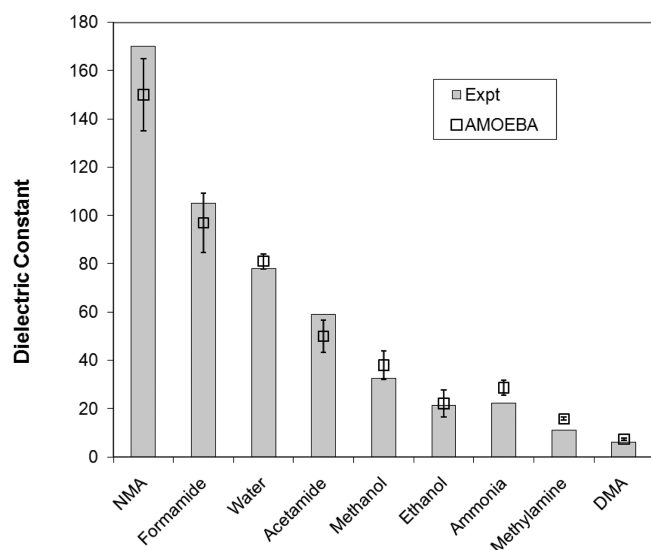


Figure 7. Dielectric constants from AMOEBA liquid MD simulations. The squares with vertical error bars are values computed from MD simulations. The filled bars are experimental values.

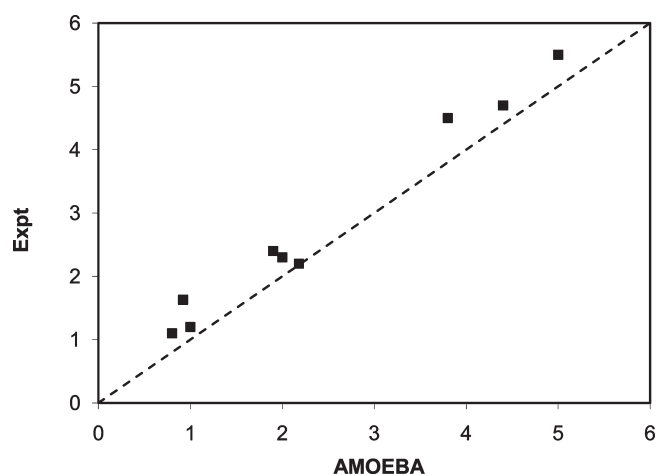


Figure 8. Comparison of diffusion constants from experimental measurement and MD simulations using AMOEBA. Diffusion constant data ($\times 10^{-9}$ m/s): dimethylformamide, $D_{\text{sim}} = 0.92$, $D_{\text{exptl}} = 1.63$, ref 186; ammonia, $D_{\text{sim}} = 5.0$, $D_{\text{exptl}} = 5.5$, ref 188; trimethylamine, $D_{\text{sim}} = 4.4$, $D_{\text{exptl}} = 4.7$ (273K), ref 189; methylamine, $D_{\text{sim}} = 3.8$, $D_{\text{exptl}} = 4.5$ interpolated from ref 189; methanol, $D_{\text{sim}} = 1.9$, $D_{\text{exptl}} = 2.4$, ref 190.

ability to reproduce the dielectric constant in the series of TIPxP models. Nonetheless, the adoption of five sites in TIPSP clearly has an effect on electrostatic interactions, as reflected in the water dimer energy surface and an increased tendency to form tetrahedral structure in the bulk.⁵⁴ Both the polarizable AMOEBA and the Drude oscillator model³³ can accurately reproduce the dielectric constant for water. Formamide has a high dielectric constant of 105. An early calculation using OPLS reported a value of 59 for formamide, while for DMF the computed value of 32 was in reasonable agreement with the experiment value of 37.¹¹⁸ In the current study, the dielectric constant of formamide is slightly underestimated by AMOEBA at 98. The dielectric constant given by the recent CHARMM Drude oscillator model was somewhat too low as well. It was suggested that the static

Table 10. Static Dielectric Constant and Self-Diffusion Coefficient ($\times 10^{-9}$ m² s⁻¹)^a

	T (K)	dielectric constant		self-diffusion		
		exptl.	AMOEBAsim	exptl.	AMOEBAsim	
water	298.2	78.4 ^b	81.0 (3.1)	2.3 ^h	2.1	
formamide	298.2	105.0 ^c	84 (293 K) ^d	97.8 (12.3)		
DMF	298.2			1.6 ⁱ	0.9	
NMA	308.2	170.0 ^d	153 (15.0)	1.2 ^j	1.0	
ammonia	240.0	22.0 ^d	28.6 (3.0)	5.5 ^k	5.0	
methylamine	266.9	10.5 ^e	16.7 (215 K) ^b	15.8 (0.6)	4.5 ^l	3.8
dimethylamine	280.0	6.0 ^d	7.3 (0.5)			
trimethylamine	276.0	2.4 ^b	1.9 (0.4)	4.7 ^l	4.4	
methanol	298.2	33.0 ^b	38.0 (5.8)	2.4 ^m	1.9	
ethanol	298.2	24.3 ^{f/g}	22.1 (5.6)	1.1 ^h	0.8	
acetamide	373.2	59 (355 K) ^d	52.4 (6.6)			
benzene	298.2	2.3 ^b	1.1 (0.5)	2.2 ⁿ	2.2	

^a The uncertainty of the calculated static dielectric constant is given in the parentheses. The uncertainty in self-diffusion constants is less than 0.1×10^{-9} m² s⁻¹. ^b Ref 151. ^c Ref 151. ^d Ref 181. ^e Ref 182. ^f Ref 183. ^g Ref 184. ^h Ref 185. ⁱ Ref 186. ^j Ref 187. ^k Refs 186 and 188. ^l Ref 189. ^m Ref 190. ⁿ Ref 191.

dielectric constant of NMA has a strong dependence on its average dipole moment, and a 0.2 D drop in the dipole moment lowered the dielectric constant by 30.³⁴ However, the induced dipole-based PIPF model overestimated the dielectric constant for NMA by 50% even though its liquid molecular dipole moment (5.0 D) is lower than that of a Drude model (5.7 D). The average NMA dipole according to AMOEBA is 5.5 D. Therefore, the dependence of the dielectric constant on the molecular dipole moment may only hold for a given, specific model. Dielectric constants of small alcohols are generally in the 20–30 range. The dielectric constant of methanol was reproduced accurately by a polarizable force field,⁴⁹ whereas a fixed charge potential underestimated the ethanol dielectric constant.¹¹⁹ These results suggest that it is difficult for the classical models to capture the static dielectric constant without explicit incorporation of polarization effects.

For self-diffusion coefficients, there seems to be a systematic underestimation by the AMOEBA model, although the errors for water, ethanol, NMA, TMA, and benzene are insignificant. The polarizable Drude oscillator model also reported reasonable diffusion coefficients for benzene and toluene.³⁵ It was noticed by the early developers of polarizable force fields⁴³ that polarization slowed diffusion in neat liquids significantly compared to the fixed charge counterparts. For water, most of the fixed charge models overestimate the diffusion coefficient by as much as a factor of 2,¹²⁰ with TIPSP and SPC/E¹¹⁶ being notable exceptions. On the other hand, the diffusion coefficients given by the CHARMM FQ model are very similar to fixed charge CHARMM22, with random errors in both directions.²⁹ Radial distribution functions for methanol and ammonia are compared to those derived from neutron scattering experiments in Figure 9–11. AMOEBA gives a more dominant first peak in the O–H RDF for liquid methanol, corresponding to the hydrogen bonding of the hydroxyl group, than that inferred from the experiment.¹²¹ A previous CPMD study has also suggested a similar peak height at about 3.3.¹²² For weakly hydrogen-bonded ammonia, the “experimental” radial distribution function¹²³ given

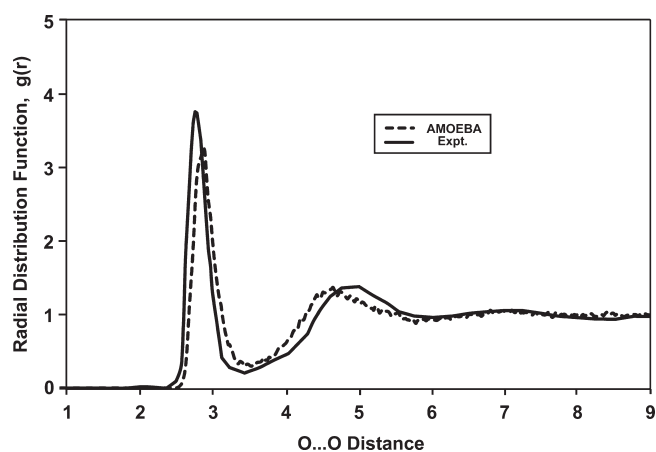


Figure 9. Radial distribution, $g(r)$, for oxygen–oxygen atom pairs in liquid methanol at 298 K.

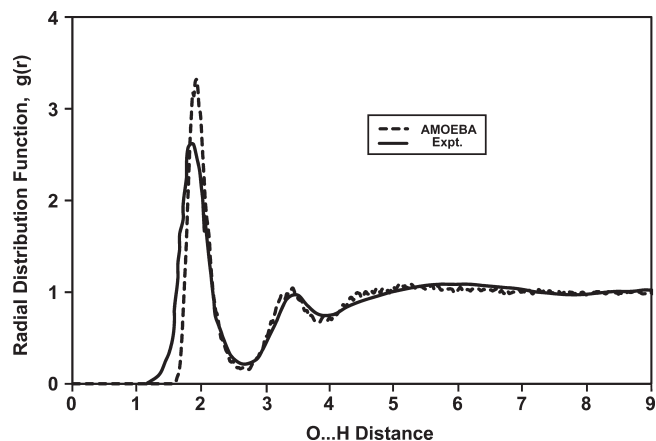


Figure 10. Radial distribution, $g(r)$, for oxygen–hydrogen atom pairs in liquid methanol at 298 K.

in Figure 11 was derived for molecular centers from X-ray scattering assuming spherical symmetry. A more recent neutron diffraction experiment¹²⁴ and first principle calculations¹²⁵ have argued that the shoulder at 3.7 Å in the early X-ray results may be artifactual, in agreement with our simulation.

Crystal Structures. The crystal structures of formamide, acetamide, acetic acid, imidazole, and 1H-indole-3-carboxaldehyde have been examined using the AMOEBA potential. Crystal models were constructed from experimental fractional coordinates and subjected to full geometry optimization of the system energy by varying both atomic coordinates and cell parameters (i.e., all cell lengths and angles). Since the unit cells of these crystals are fairly small, replicated supercells were computed to allow use of particle mesh Ewald for long-range electrostatics. After full optimization, the atomic coordinates deviated from the experimental crystal by at most 0.3 Å in all cases (Table 11). In general, the overall cell volume shrank slightly, as expected for energy minimization. Molecular dynamics simulations of these and other crystals at experimental temperatures are underway and will be reported in due course.

Hydration Free Energy. Solvation plays a critical role in many chemical and biological processes. Accurate knowledge of solvation energetics is needed as part of the calculation of absolute

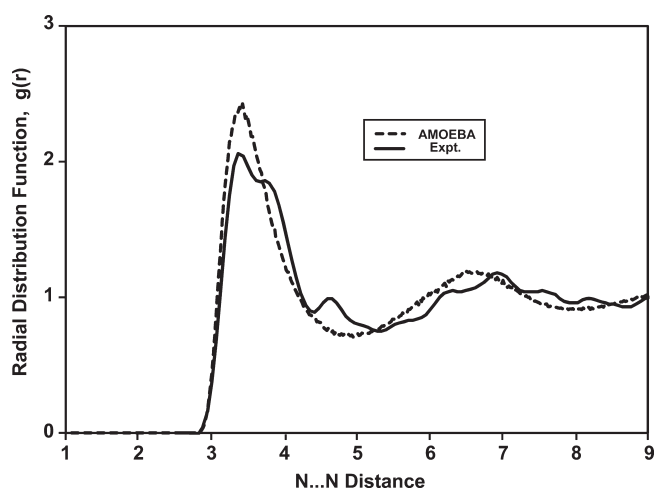


Figure 11. Radial distribution, $g(r)$, for nitrogen–nitrogen atom pairs in liquid ammonia at 277 K.

association energies, for example, the binding of ligands to proteins. There is an extensive history of estimating the solvation free energy for small organics, protein side chain analogs, etc. using various force fields and water models. Vianna and Mark calculated the hydration free energies of 18 small molecules using a GROMOS96 force field¹²⁶ and SPC water model.¹²⁷ The average error was 2.8 kcal/mol using the original GROMOS partial charges, and it was suggested that the error might be reduced to 1 kcal/mol if the charge values were increased by 10%. In similar work by Maccallum and Tieleman using the OPLS-AA force field for solutes, an average unsigned error of 1.1 kcal/mol was reported for OPLS-AA in TIP4P water, 1.2 for OPLS-AA in SPC water, and 2.1 kcal/mol for GROMOS96 in SPC water.¹²⁸ Later, the GROMOS force field was optimized to reproduce the solvation free energies in water and cyclohexane, resulting in a much smaller error (0.2 kcal/mol).¹²⁹ However, this last study required the solutes to adopt different atomic charges in water and cyclohexane. Recently, effort has been devoted to increasing precision in hydration free energy calculations and optimizing the force field charges to capture solvation free energy more accurately. Shirts and co-workers¹³⁰ showed that it is possible to reduce the statistical uncertainty in the calculated hydration free energy to below 0.05 kcal/mol, and exhaustive sampling of various parameters was achieved using the folding@home computing resource.¹³¹ Subsequently, the hydration free energy of 15 amino acid side chain analogs was determined using OPLS-AA and the above-mentioned water models plus SPC/E, TIP4P-EW,¹³² and TIP3P-MOD.^{133,134} The TIP3P-MOD, a modified TIP3P model to improve the solvation free energy of methane, gave the most accurate hydration free energy values with a RMSE of 0.51 kcal/mol. It is interesting to note that the TIP4P-EW model that yields the best overall pure water properties led to the worst hydration energies among all water models tested. Further modification of TIP3P vdW parameters in the spirit of TIP3P-MOD has been able to optimize the solvation energy of all 15 compounds to a RMSE of 0.39 kcal/mol. However, it should be cautioned that changes in the vdW parameters have profound effects on bulk water properties. While the heat of vaporization and density may remain reasonable, the structure of water (e.g., radial distribution function for O...O and O...H pair distances) is very sensitive to the vdW parameters.

Table 11. Comparison of Experimental and AMOEBA-Optimized Crystal Structures and Cell Parameters of Organic Molecules^a

		struct RMSE	cell	<i>a</i>	<i>b</i>	<i>c</i>	α	β	γ	ref
formamide	exptl (90K)		3 × 1 × 2	10.812	9.041	13.988	90	100.5	90	192
	calcd	0.3		10.643	9.340	13.497	90	104.3	90	
acetamide	exptl (23K)		1 × 1 × 1	11.513	11.513	12.883	90	90	120	193
	calcd	0.1		11.564	11.564	12.289	90	90	120	
acetic acid	exptl (83K)		1 × 3 × 2	13.214	11.772	11.532	90	90	90	194and195
	calcd	0.1		13.424	11.573	11.352	90	90	90	
imidazole	exptl (293K)		2 × 3 × 2	15.464	16.374	19.558	90	117.3	90	196
	calcd	0.3		14.756	15.718	20.100	90	117.2	90	
1H-indole 3-carbox-aldehyde	exptl (295K)		1 × 2 × 2	14.145	11.664	17.428	90	90	90	197
	calcd	0.3		14.458	11.964	16.683	90	90	90	

^aThe cell lengths are in Ångstroms, and angles are in degrees.

Table 12. Hydration Free Energies of Small Molecules (kcal/mol)^a

molecule	AMOEBA	exptl
methane	1.73 (0.13)	1.99 ^b
ethane	1.73 (0.15)	1.83 ^b
propane	1.69 (0.17)	1.96 ^b
<i>n</i> -butane	1.11 (0.21)	2.08 ^b
methanol	-4.79 (0.23)	-5.11 ^b
ethanol	-4.69 (0.25)	-5.00 ^b
propanol	-4.85 (0.27)	-4.83 ^b
isopropanol	-4.21 (0.34)	-4.76 ^b
phenol	-5.05 (0.28)	-6.62 ^b
<i>p</i> -cresol	-5.60 (0.31)	-6.14 ^b
methylether	-2.22 (0.38)	-1.90 ^b
benzene	-1.23 (0.23)	-0.87 ^b
toluene	-1.53 (0.25)	-0.89 ^b
ethylbenzene	-0.80 (0.28)	-0.80 ^b
methylamine	-5.46 (0.25)	-4.56 ^b
ethylamine	-4.33 (0.24)	-4.50 ^b
dimethylamine	-3.04 (0.26)	-4.29 ^b
trimethylamine	-2.09 (0.24)	-3.24 ^b
imidazole	-10.25 (0.30)	-9.63 ^c
<i>N</i> -methylacetamide	-8.66 (0.30)	-10.00 ^d
acetic acid	-5.63 (0.20)	-6.70 ^b
hydrogen sulfide	-0.41 (0.17)	-0.44 ^b
methylsulfide	-1.43 (0.27)	-1.24 ^b
ethylsulfide	-1.74 (0.24)	-1.30 ^b
dimethylsulfide	-1.85 (0.22)	-1.54 ^b
methylethylsulfide	-1.98 (0.32)	-1.50 ^e
water	-5.86 (0.19)	-6.32 ^e

^aStatistical uncertainties of AMOEBA calculations are given in parentheses. ^bRef 198. ^cRef 199. ^dRef 200. ^eRef 201.

More recently Mobley et al. took a different approach, investigating the effect of solute charges on hydration free energy.¹³⁵ Among the protocols they tested, RESP charges from HF/6-31G* performed the best, with a RMSE of 1.04 kcal/mol in hydration free energy of 44 compounds, closely followed (RMSE = 1.10 kcal/mol) by semiempirical charges from an AM1-BCC method tuned to reproduce HF/6-31G* charges.^{136,137} Recent calculations using Amber GAFF parameters for 504 neutral

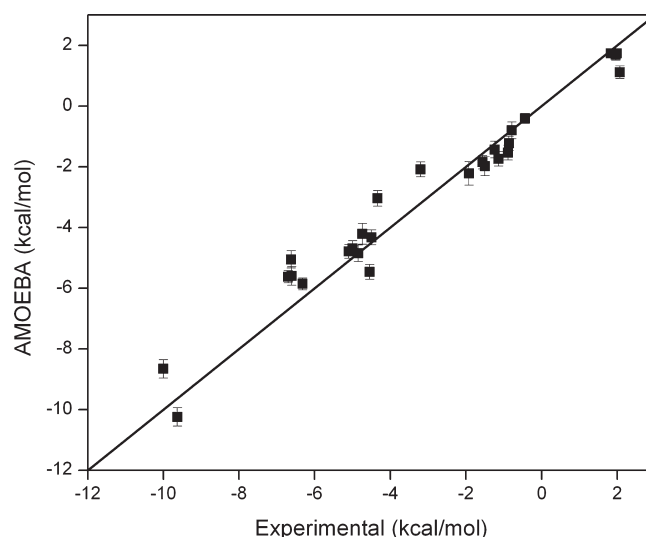


Figure 12. Comparison of solvation free energies of 27 small molecules calculated with the AMOEBA force field with the experimental values. Signed average error = -0.11 kcal/mol; unsigned average error = 0.56 kcal/mol; RMSE = 0.69 kcal/mol.

molecules reported a RMSE of 1.24 kcal/mol and a correlation of 0.89 between simulation and experimental values.¹³⁸

In the current study, we have computed the hydration free energy of 27 compounds as a validation of the AMOEBA force field. None of the hydration free energy data was incorporated into the parametrization process. Results are listed in Table 12, and the correlation with experimental data is plotted in Figure 12. For this small set of compounds, the RMSE between AMOEBA and experimental results is 0.69 kcal/mol. The average signed error is 0.11, and the average unsigned error is 0.56 kcal/mol. The correlation (R^2) between the calculated and experimental value is 0.96 with a slope of 1.09. The largest error was observed for phenol, at 1.57 kcal/mol. Five out of 27 compounds had a deviation from experimental results greater than 1 kcal/mol. There is no obvious correlation between the error in hydration free energy and errors in gas-phase dimer energy or neat liquid heat of vaporization. The errors for methyl, dimethyl, and trimethyl amine are all about 1 kcal/mol. The densities of both dimethyl- and trimethylamine are higher than experimental results, while the density of methylamine is underestimated, as

indicated by the average pressure from NVT simulations reported in Table 8. In retrospect, it seems likely that the methyl vdW parameters, shared by all three amines, are not fully optimized for the condensed phase. The solvation energy RMSE for the other 24 compounds is 0.45 kcal/mol, indicating that there remains room for improvement in the amine solvation energies.

We note that AMOEBA gives a RMSE of 0.23 kcal/mol for liquid heat of vaporization and 0.38 kcal/mol for dimer energy in the gas phase, two properties that were actively utilized during parameter optimization. Thus, the ideal RMSE target of the calculated hydration free energy should probably lie below 0.5 kcal/mol. The error may come from various sources. The three main contributions to intermolecular interaction in the AMOEBA model arise from permanent electrostatics by atomic multipoles, polarization via atomic polarizability, and vdW interactions. It is possible that the level of *ab initio* theory and basis set used to derive the atomic multipoles is not sufficient. The buffered 14–7 vdW potential, and particularly the effect of a combining rule on heteroatomic vdW interactions in differing environments, is a likely impediment to improved accuracy.¹³⁹

CONCLUSIONS

A polarizable point multipole potential has been developed for a range of common small organic molecules. Molecular electrostatics is represented by atomic multipole moments through the quadrupole, while polarization effects are treated via classical induced dipole interactions. The permanent atomic multipoles are derived from *ab initio* theoretical calculations at the MP2/6-311G(1d,1p) and MP2/aug-cc-pVTZ levels. Dipole polarization is modeled empirically with a damped, interactive atomic dipole scheme and using a small set of highly transferable atomic polarizabilities. The vdW parameters are derived via simultaneous fitting to gas-phase dimer calculations and liquid thermodynamic properties such as the density and heat of vaporization. Sharing atom types within and across families of organic molecules ensures transferability of parameters. A number of other gas phase and condensed phase properties were computed for use in validation, including stable homo- and heterodimer energies and structures, liquid diffusion and dielectric constants, radial distribution functions, molecular crystal structures, and solvation free energies in water. Overall, satisfying agreement between the polarizable potential, *ab initio*, and experimental results has been achieved in both the gas and condensed phases. The improvement in energy and density of homogeneous liquids is modest compared to a well-tuned fixed charge force field such as OPLS-AA, but the introduction of polarization and atomic multipoles significantly improves the ability of AMOEBA to describe details of molecular interactions across different environments. Current and prior results using this polarizable force field³⁴ suggest that the inclusion of induction effects is crucial for capturing diffusion and dielectric properties. The hydration free energy of 27 organic compounds computed using the current parameters verifies the general transferability of the model (RMSE = 0.69 kcal/mol), but further improvement is likely still possible.

A realistic physical model is critical for an accurate and transferable empirical force field. Obtaining consistent parameters for such a model presents an immense challenge. On the basis of the lessons learned in the current study and recent work by others on polarizable force fields, we expect that the overall performance of the AMOEBA model can be further refined.

Nowadays, accurate QM calculations can be performed routinely on small- to moderate-sized organics. The DMA procedure combined with potential fitting allows us to utilize high level *ab initio* calculations directly in AMOEBA parametrization. While our current feeling is that the AMOEBA permanent electrostatics are sufficient to construct a highly accurate force field, there are some indications that the polarization model can be improved in comparison to rigorous quantum results.¹⁴⁰ Given the increasing availability of computing resources, it should be possible to systematically optimize vdW parameters to reproduce neat liquid properties and transfer free energies simultaneously. The inclusion of Tang–Tonnesen damping of dispersion interactions at short-range¹⁴¹ is formally analogous to Thole damping of polarization effects and may be important for applications such as crystal structure prediction.¹⁴² An important omission in most current force fields, including AMOEBA, is explicit coupling of electrostatics to the valence parameters. Simple schemes have been proposed to include this coupling,¹⁴³ and it is known to play a role in bond angle deformation in liquid water,¹⁴⁴ pyramidalization at amide nitrogen atoms,¹⁴⁵ and other important structural features. Future studies should also move beyond calculation of hydration free energy to include an examination of free energies of transfer, solvation structure around the solutes, and additional dynamic properties using AMOEBA and alternative polarizable force fields.

APPENDIX

Polarization Energy Gradient. A derivation of the gradient of the AMOEBA polarization energy is provided below. It is convenient to express the system energy via super-matrices:

$$U_{\text{ele}}^{\text{perm}} = \frac{1}{2} M^T T M$$

$$\text{with } M^T = [M_1 \dots M_i \dots M_N]$$

$$\text{and } T = \begin{bmatrix} 0 & T_{12} & \dots & T_{1n} \\ T_{21} & 0 & \dots & T_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n1} & T_{n2} & \dots & 0 \end{bmatrix} \quad (\text{A1})$$

where M_i is the transposed permanent multipole vector at site i , and T_{ij} is the interaction matrix between site i and j :

$$T_{ij} = \begin{bmatrix} 1 & \frac{\partial}{\partial x_j} & \frac{\partial}{\partial y_j} & \frac{\partial}{\partial z_j} & \dots \\ \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial z_j} & \dots \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial z_j} & \dots \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \left(\frac{1}{r_{ji}} \right) \quad (\text{A2})$$

Rewriting eq 4 in terms of a super-matrix yields

$$(\alpha^{-1} - T^{11}) \mu^{\text{ind}} = T^1 M = E \quad (\text{A3})$$

Here, μ^{ind} is a vector of length $3N$, where N is the number of polarizable sites, $\mu^{\text{ind}} = [\mu_{1x}, \mu_{1y}, \mu_{1z}, \dots, \mu_{Nx}]^T$. α^{-1} is a $3N \times 3N$ matrix with α_{1x}^{-1} , α_{1y}^{-1} , etc. as diagonal components and all

off-diagonal components equal to zero. T^1 is a super matrix with elements corresponding to the field tensor T_{ij}^1 in eq 4 (i.e., $3N \times 13N$).

Now, we define $C = \alpha^{-1} - T^{11}$ and note that it is a symmetric matrix such that $C^T = C$. The induction energy is then defined by the product of the induced dipole with the permanent field

$$U_{\text{ele}}^{\text{ind}} = -\frac{1}{2} (\mu^{\text{ind}})^T E = -\frac{1}{2} E^T C^{-1} E \quad (\text{A4})$$

Subsequently, the energy gradient on site k is given by

$$\frac{\partial U_{\text{ele}}^{\text{ind}}}{\partial x_k} = -\frac{1}{2} \left(\frac{\partial E^T}{\partial x_k} C^{-1} E + E^T \frac{\partial C^{-1}}{\partial x_k} E + E^T C^{-1} \frac{\partial E}{\partial x_k} \right),$$

$$k = 1, 2, 3, \dots, 3N \quad (\text{A5})$$

The gradient on the left is a $3N$ vector in the above equation. Given $C^{-1}C = I$, i.e., $(\partial C^{-1}/\partial x_k)C + C^{-1}(\partial C/\partial x_k) = 0$, we can simplify to obtain

$$\begin{aligned} \frac{\partial U_{\text{ele}}^{\text{ind}}}{\partial x_k} &= -\frac{1}{2} \frac{\partial E^T}{\partial x_k} \mu^{\text{ind}} + \frac{1}{2} E^T C^{-1} \frac{\partial C}{\partial x_k} C^{-1} E - \frac{1}{2} (\mu^{\text{ind}})^T \frac{\partial E}{\partial x_k} \\ &= -(\mu^{\text{ind}})^T \frac{\partial T^1 M}{\partial x_k} - \frac{1}{2} (\mu^{\text{ind}})^T \frac{\partial T^{11}}{\partial x_k} \mu^{\text{ind}} \end{aligned} \quad (\text{A6})$$

Given the total multipoles at each site as $M^t = M + M^{\text{ind}}$, the net force at the site becomes

$$\frac{\partial U_{\text{ele}}^t}{\partial x_k} = -\frac{1}{2} (M^t)^T \frac{\partial T^1}{\partial x_k} M^t - (M^t)^T T^1 \frac{\partial R}{\partial x_k} M \quad (\text{A7})$$

where the factor 1/2 takes care of the redundancy due to the inclusion of both ij and ji in the above summation. While permanent multipoles in the local frame are invariant parameters, the permanent multipole moments, M , in the global frame are a function of the local frame rotation matrix, leading to additional chain rule terms related to the rotational force, i.e., a torque. The exact formula for the force and torque components can be easily obtained by comparing the above to the permanent–permanent terms¹⁸ and keeping in mind the additional factor of 1/2. Note that the torque term does not include an induced–induced contribution since induced dipoles are always defined in the global frame.

When the field “ E ” in the induction energy is different from the “ E ” that produces the induced dipoles, the gradient formula requires further modification. The induction energy is then given by

$$U_{\text{ele}}^{\text{ind}} = -\frac{1}{2} (\mu_{\text{d}}^{\text{ind}})^T E_{\text{p}} = -\frac{1}{2} E_{\text{d}}^t C^{-1} E_{\text{p}} \quad (\text{A8})$$

where E_{p} is the field actually used in the polarization energy calculation, and E_{d} is the “direct” field due to permanent multipoles responsible for polarization. The difference between the two subscripts, d and p, results from differing local interaction scaling. In traditional molecular mechanics, short range nonbonded interactions between bonded atoms are generally neglected. In the current model, the interaction energy between E_{p} and induced dipoles is ignored for 1–2, 1–3, etc. bonded pairs, as these effects are implicitly included in bond and angle terms, while we recall that intramolecular “direct” polarization occurs between polarization groups.

The gradient of the above energy becomes

$$\frac{\partial U^{\text{ind}}}{\partial x_k} = -\frac{1}{2} E_{\text{d}}^T C^{-1} E_{\text{p}} \quad (\text{A9})$$

$$\frac{\partial U^{\text{ind}}}{\partial x_k} = -\frac{1}{2} \left(\frac{\partial E_{\text{d}}^T}{\partial x_k} C^{-1} E_{\text{p}} + E_{\text{d}}^T \frac{\partial C^{-1}}{\partial x_k} E_{\text{p}} + E_{\text{d}}^T C^{-1} \frac{\partial E_{\text{p}}}{\partial x_k} \right) \quad (\text{A10})$$

Now define an intermediate quantity, $\mu_{\text{p}}^{\text{ind}} = C^{-1} E_{\text{p}}$. Recall that C is invariant with respect to local interaction scaling since “mutual” induction always occurs between every atom pair.

$$\begin{aligned} \frac{\partial U^{\text{ind}}}{\partial x_k} &= -\frac{1}{2} \left(\frac{\partial E_{\text{d}}^T}{\partial x_k} \mu_{\text{p}}^{\text{ind}} - E_{\text{d}}^T C^{-1} \frac{\partial C}{\partial x_k} C^{-1} E_{\text{p}} + (\mu_{\text{d}}^{\text{ind}})^T \frac{\partial E_{\text{p}}}{\partial x_k} \right) \\ &= -\frac{1}{2} \left[\frac{\partial E_{\text{d}}^T}{\partial x_k} \mu_{\text{p}}^{\text{ind}} + (\mu_{\text{d}}^{\text{ind}})^T \frac{\partial E_{\text{p}}}{\partial x_k} \right] - \frac{1}{2} (\mu_{\text{d}}^{\text{ind}})^T \frac{\partial T^{11}}{\partial x_k} \mu_{\text{p}}^{\text{ind}} \\ &= -\frac{1}{2} \left[\frac{\partial T^1}{\partial x_k} \mu_{\text{p}}^{\text{ind}} + (\mu_{\text{d}}^{\text{ind}})^T \frac{\partial T^1 M}{\partial x_k} \right] - \frac{1}{2} (\mu_{\text{d}}^{\text{ind}})^T \frac{\partial T^{11}}{\partial x_k} \mu_{\text{p}}^{\text{ind}} \end{aligned} \quad (\text{A11})$$

In cases where there is no intermolecular polarization (e.g., water molecule), $\mu_{\text{p}}^{\text{ind}}$ equals $\mu_{\text{d}}^{\text{ind}}$, and the above equation reduces to eq A6. In practice, the two sets of μ are converged simultaneously, as the only difference between the two is the scaling of real-space local interactions.

Upon comparing eqs A6 and A11, it can be seen that within induced–permanent terms the induced dipole on site i (where force is computed) is replaced by $1/2 (\mu_{\text{d}}^{\text{ind}} + \mu_{\text{p}}^{\text{ind}})$ and the induced–induced term for a given pair interaction, $(\mu_{\text{d}}^{\text{ind}})^T (\partial T^{11}/\partial x_k) \mu_{\text{p}}^{\text{ind}}$, is replaced by $1/2 ((\mu_{\text{d}}^{\text{ind}})^T (\partial T^{11}/\partial x_k) \mu_{\text{p}}^{\text{ind}} + (\mu_{\text{p}}^{\text{ind}})^T (\partial T^{11}/\partial x_k) \mu_{\text{d}}^{\text{ind}})$. Throughout the energy, force, torque, and virial terms, similar substitutions can be made for induced–permanent terms and induced–induced terms in the Ewald formulation. The algorithms for computing the permanent and induced force, torque, and virial, both for pairwise non-periodic systems and when using PME, have been implemented and verified numerically in the TINKER and AMBER software packages.

■ ASSOCIATED CONTENT

S Supporting Information. Detailed protocol outlining a step-by-step parametrization procedure for determination of AMOEBA values for new organic molecules is provided. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ponder@dasher.wustl.edu (J.W.P.); pren@mail.utexas.edu (P.R.).

■ ACKNOWLEDGMENT

P.R. acknowledges support by the National Institute of General Medical Sciences (R01 GM079686) and Robert A. Welch Foundation (F-1691). J.W.P. acknowledges support from the National Science Foundation (Award 0535675) and the National Institutes of Health (R01 GM069553). The AMOEBA parameters for organic molecules are available as part of the

TINKER modeling package, which can be obtained from <http://dasher.wustl.edu/tinker/>.

REFERENCES

- (1) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- (2) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (3) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (4) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (5) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (6) Silberstein, L. *Philos. Mag. Ser. 6* **1917**, *33*, 92–128.
- (7) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (8) Vesely, F. J. *J. Comput. Phys.* **1977**, *24*, 361–371.
- (9) Sprik, M. *J. Phys. Chem.* **1991**, *95*, 2283–2291.
- (10) Dang, L. X.; Chang, T. M. *J. Chem. Phys.* **1997**, *106*, 8149–8159.
- (11) Brdarski, S.; Astrand, P. O.; Karlstrom, G. *Theor. Chem. Acc.* **2000**, *105*, 7–14.
- (12) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730–4737.
- (13) Burnham, C. J.; Li, J. C.; Xantheas, S. S.; Leslie, M. J. *J. Chem. Phys.* **1999**, *110*, 4566–4581.
- (14) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237–2251.
- (15) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- (16) Burnham, C. J.; Xantheas, S. S. *J. Chem. Phys.* **2002**, *116*, 1479–1492.
- (17) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515–1531.
- (18) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- (19) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621–627.
- (20) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2004**, *108*, 13427–13437.
- (21) Jiao, D.; King, C.; Grossfield, A.; Darden, T. A.; Ren, P. Y. *J. Phys. Chem. B* **2006**, *110*, 18553–18559.
- (22) Xie, W. S.; Pu, J. Z.; MacKerell, A. D.; Gao, J. L. *J. Chem. Theory Comput.* **2007**, *3*, 1878–1889.
- (23) Rappé, A. K.; Goddard, W. A., III. *J. Phys. Chem.* **1991**, *95*, 3358–3363.
- (24) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141–6156.
- (25) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. J. *J. Mol. Liq.* **1995**, *65–6*, 31–40.
- (26) Banks, J. L.; Kaminski, G. A.; Zhou, R. H.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741–754.
- (27) Ando, K. *J. Chem. Phys.* **2001**, *115*, 5228–5237.
- (28) Yoshii, N.; Miyauchi, R.; Miura, S.; Okazaki, S. *Chem. Phys. Lett.* **2000**, *317*, 414–420.
- (29) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1–15.
- (30) Patel, S.; MacKerell, A. D.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1504–1514.
- (31) van Maaren, P. J.; van der Spoel, D. *J. Phys. Chem. B* **2001**, *105*, 2618–2626.
- (32) Yu, H. B.; Hansson, T.; van Gunsteren, W. F. *J. Chem. Phys.* **2003**, *118*, 221–234.
- (33) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185–5197.
- (34) Harder, E.; Anisimov, V. M.; Whitfield, T. W.; MacKerell, A. D.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509–3521.
- (35) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D. *J. Phys. Chem. B* **2007**, *111*, 2873–2885.
- (36) Rick, S. W.; Stuart, S. J. *Rev. Comp. Ch.* **2002**, *18*, 89–146.
- (37) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.
- (38) Cieplak, P.; Dupradeau, F. Y.; Duan, Y.; Wang, J. M. *J. Phys., Condens. Mater.* **2009**, *21*, 333101.
- (39) Illingworth, C. J.; Domene, C. *Proc. R. Soc. London, Ser. A* **2009**, *465*, 1701–1716.
- (40) Lopes, P. E. M.; Lamoureux, G.; Mackerell, A. D. *J. Comput. Chem.* **2009**, *30*, 1821–1838.
- (41) Masia, M.; Probst, M.; Rey, R. *J. Chem. Phys.* **2004**, *121*, 7362–7378.
- (42) Masia, M.; Probst, M.; Rey, R. *J. Chem. Phys.* **2005**, *123*, 164505.
- (43) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208–6219.
- (44) Gao, J. L.; Pavelites, J. J.; Habibollahzadeh, D. *J. Phys. Chem.* **1996**, *100*, 2689–2697.
- (45) Cabaleiro-Lago, E. M.; Rios, M. A. *J. Chem. Phys.* **1998**, *108*, 3598–3607.
- (46) Hermida-Ramon, J. M.; Rios, M. A. *J. Phys. Chem. A* **1998**, *102*, 10818–10827.
- (47) Qian, W. L.; Krimm, S. *J. Phys. Chem. A* **2001**, *105*, 5046–5053.
- (48) Mannfors, B.; Mirkin, N. G.; Palmo, K.; Krimm, S. *J. Comput. Chem.* **2001**, *22*, 1933–1943.
- (49) Yu, H. B.; Geerke, D. P.; Liu, H. Y.; van Gunsteren, W. F. *J. Comput. Chem.* **2006**, *27*, 1494–1504.
- (50) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587–1597.
- (51) Williams, D. E. *J. Comput. Chem.* **1988**, *9*, 745–763.
- (52) Dykstra, C. E. *Chem. Rev.* **1993**, *93*, 2339–2353.
- (53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (54) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.
- (55) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, U. K., 1996.
- (56) Poplelier, P. L. A.; Joubert, L.; Kosov, D. S. *J. Phys. Chem. A* **2001**, *105*, 8254–8261.
- (57) Buckingham, A. D.; Fowler, P. W. *J. Chem. Phys.* **1983**, *79*, 6426–6428.
- (58) Buckingham, A. D.; Fowler, P. W. *Can. J. Chem.* **1985**, *63*, 2018–2025.
- (59) Golubkov, P. A.; Ren, P. *J. Chem. Phys.* **2006**, *125*, 064103.
- (60) Golubkov, P. A.; Wu, J. C.; Ren, P. Y. *J. Phys. Chem. Chem. Phys.* **2008**, *10*, 2050–2057.
- (61) Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (62) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341–350.
- (63) van Duijnen, P. T.; Swart, M. *J. Phys. Chem. A* **1998**, *102*, 2399–2407.
- (64) Piquemal, J.-P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 054511.
- (65) Price, S. L.; Faerman, C. H.; Murray, C. W. *J. Comput. Chem.* **1991**, *12*, 1187–1197.
- (66) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233–239.
- (67) Young, D. M. *Iterative Solution of Large Linear Systems*; Academic Press: New York, 1971.
- (68) Smith, W. *CCP5 Newsletter* **1998**, *46*, 18–30.
- (69) Sagui, C.; Darden, T.; Pedersen, L. G. *J. Chem. Phys.* **2004**, *120*, 73–87.
- (70) Kong, Y. Ph.D. thesis, Molecular Biophysics, Washington University Medical School, St. Louis, Missouri, 1997.
- (71) Kolafa, J. *J. Chem. Phys.* **2005**, *122*, 164105.

- (72) Sala, J.; Guardia, E.; Masia, M. *J. Chem. Phys.* **2010**, *133*, 234101.
- (73) van Belle, D.; Wodak, S. J. *Comput. Phys. Commun.* **1995**, *91*, 253–262.
- (74) Harder, E.; Kim, B.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2005**, *1*, 169–180.
- (75) Souaille, M.; Loirat, H.; Borgis, D.; Gaigeot, M. P. *Comput. Phys. Commun.* **2009**, *180*, 276–301.
- (76) Darden, T. A.; Toukmaji, A.; Pedersen, L. G. *J. Chim. Phys. PCB* **1997**, *94*, 1346–1364.
- (77) Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol.* **1999**, *28*, 155–179.
- (78) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian Inc.: Wallingford, CT, 2003.
- (79) Stone, A. J. *GDMA*; Cambridge University Technical Services: Cambridge, England, 1998.
- (80) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.
- (81) Shi, Y.; Wu, C.; Ponder, J. W.; Ren, P. *J. Comput. Chem.* **2010**, *32*, 967–977.
- (82) Ponder, J. W. *TINKER Molecular Modeling Package*, V5.1; Washington University Medical School: St. Louis, MO, 2010.
- (83) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (84) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6290–6295.
- (85) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (86) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. *J. Chem. Theory Comput.* **2011**, *7*, 790–797.
- (87) Kollman, P. A. *J. Am. Chem. Soc.* **1971**, *94*, 1837–1842.
- (88) Reiher, W. E. Ph.D. Thesis, Dept. of Chemistry, Harvard University, Cambridge, MA, 1985.
- (89) Lü, J. H.; Allinger, N. L. *J. Comput. Chem.* **1998**, *19*, 1001–1016.
- (90) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. *Chem.—Eur. J.* **2009**, *15*, 851–855.
- (91) Steiner, T. *Angew. Chem., Int. Ed.* **2002**, *41*, 48–76.
- (92) Holt, A.; Boström, J.; Karlström, G.; Smith, R. J. *Comput. Chem.* **2010**, *31*, 1583–1591.
- (93) Cieplak, P.; Caldwell, J.; Kollman, P. J. *Comput. Chem.* **2001**, *22*, 1048–1057.
- (94) Baker, C. M.; Grant, G. H. *J. Chem. Theory Comput.* **2006**, *2*, 947–955.
- (95) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Am. Chem. Soc.* **2000**, *122*, 4750–4755.
- (96) Bartlett, G. J.; Choudhary, A.; Raines, R. T.; Woolfson, D. N. *Nat. Chem. Biol.* **2010**, *6*, 615–620.
- (97) Vargas, R.; Garza, J.; Friesner, R. A.; Stern, H.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2001**, *105*, 4963–4968.
- (98) Sponer, J.; Hobza, P. *J. Phys. Chem. A* **2000**, *104*, 4592–4597.
- (99) Grossfield, A.; Ren, P. Y.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.
- (100) Jurecka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365*, 89–94.
- (101) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (102) Allinger, N. L.; Li, F.; Yan, L.; Tai, J. C. *J. Comput. Chem.* **1990**, *11*, 868–895.
- (103) Boese, A. D.; Chandra, A.; Martin, J. M. L.; Marx, D. *J. Chem. Phys.* **2003**, *119*, 5965–5980.
- (104) Janeiro-Barral, P. E.; Mella, M. *J. Phys. Chem. A* **2006**, *110*, 11244–11251.
- (105) van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt, F. B. *THEOCHEM* **1982**, *89*, 185–201.
- (106) Kukolich, S. G. *Chem. Phys. Lett.* **1970**, *5*, 401–404.
- (107) Kukolich, S. G.; Casleton, K. H. *Chem. Phys. Lett.* **1973**, *18*, 408–410.
- (108) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656–10668.
- (109) Pitonak, M.; Neogrady, P.; Rezac, J.; Jurecka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829–1834.
- (110) Dinadayalane, T. C.; Leszczynski, J. *Struct. Chem.* **2009**, *20*, 11–20.
- (111) Dinadayalane, T. C.; Leszczynski, J. *J. Chem. Phys.* **2009**, *130*, 081101.
- (112) Sun, H. *J. Phys. Chem. B* **1998**, *102*, 7338–7364.
- (113) Sun, H.; Ren, P.; Fried, J. R. *Comput. Theor. Polym. Sci.* **1998**, *8*, 229–246.
- (114) Jorgensen, W. L.; Jenson, C. J. *Comput. Chem.* **1998**, *19*, 1179–1186.
- (115) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullmann, B., Ed.; D. Reidel Pub. Co.: Dordrecht, The Netherlands, 1981; pp 331–342.
- (116) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (117) Hohtl, P.; Boresch, S.; Bitomsky, W.; Steinhauser, O. *J. Chem. Phys.* **1998**, *109*, 4927–4937.
- (118) Essex, J. W.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 17956–17962.
- (119) Saiz, L.; Guardia, E.; Padro, J. A. *J. Chem. Phys.* **2000**, *113*, 2814–2822.
- (120) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2001**, *114*, 363–366.
- (121) Yamaguchi, T.; Hidaka, K.; Soper, A. K. *Mol. Phys.* **1999**, *97*, 603–605.
- (122) Pagliai, M.; Cardini, G.; Righini, R.; Schettino, V. *J. Chem. Phys.* **2003**, *119*, 6655–6662.
- (123) Narten, A. H. *J. Chem. Phys.* **1976**, *66*, 3117–3120.
- (124) Ricci, M. A.; Nardone, M.; Ricci, F. P.; Andreani, C.; Soper, A. K. *J. Chem. Phys.* **1995**, *102*, 7650–7655.
- (125) Hannongbua, S. *J. Chem. Phys.* **2000**, *113*, 4707–4712.
- (126) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205–1218.
- (127) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548–553.
- (128) Maccallum, J. L.; Tieleman, D. P. *J. Comput. Chem.* **2003**, *24*, 1930–1935.
- (129) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (130) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
- (131) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903–1904.
- (132) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (133) Sun, Y.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1164–1169.
- (134) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
- (135) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
- (136) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.
- (137) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- (138) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.
- (139) Al-Matar, A. K.; Rockstraw, D. A. *J. Comput. Chem.* **2004**, *25*, 660–668.

- (140) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2004**, *120*, 9903–9906.
- (141) Tang, K. T.; Tonnies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726–3741.
- (142) Mooij, W. T. M.; van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Eijck, B. P. *J. Phys. Chem. A* **1999**, *103*, 9872–9882.
- (143) Palmo, K.; Mannfors, B.; Mirkin, N. G.; Krimm, S. *Chem. Phys. Lett.* **2006**, *429*, 628–632.
- (144) Fanourgakis, G. S.; Xantheas, S. S. *J. Chem. Phys.* **2006**, *124*, 174504.
- (145) Mannfors, B.; Mirkin, N. G.; Palmo, K.; Krimm, S. *J. Phys. Chem. A* **2003**, *107*, 1825–1832.
- (146) Applequist, J.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952–2960.
- (147) Bosque, R.; Sales, J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1154–1163.
- (148) Applequist, J. *J. Phys. Chem.* **1993**, *97*, 6016–6023.
- (149) Allinger, N. L.; Fermann, J. T.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1997**, *106*, 5143–5150.
- (150) Murphy, W. F.; Fernandezsanchez, J. M.; Raghavachari, K. *J. Phys. Chem.* **1991**, *95*, 1124–1139.
- (151) *CRC Handbook of Chemistry and Physics*, 82nd ed.; Lide, D. R., Ed.; CRC Press LLC: Boca Raton, FL, 2001.
- (152) Riddick, J. A.; Bunger, W. B.; Sakano, T.; Weissberger, A. *Organic Solvents: Physical Properties and Methods of Purification*, 4th ed.; Wiley: New York, 1986.
- (153) Wagman, D. D.; Evans, W. H.; Parker, V. B.; Schumm, R. H.; Halow, I. *J. Phys. Chem. Ref. Data* **1982**, *11* (Suppl 2), 1–405.
- (154) Haar, L.; Gallagher, J. S. *J. Phys. Chem. Ref. Data* **1978**, *7*, 635–793.
- (155) Aston, J. G.; Siller, C. W.; Messerly, G. H. *J. Am. Chem. Soc.* **1937**, *59*, 1743–1751.
- (156) Felsing, W. *Ind. Eng. Chem.* **1929**, *21*, 1269–1272.
- (157) Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*, 3d ed.; McGraw-Hill: New York, 1977.
- (158) Swift, E. *J. Am. Chem. Soc.* **1942**, *64*, 115–116.
- (159) Letcher, T. M. *J. Chem. Thermodyn.* **1972**, *5*, 159–173.
- (160) Aston, J. G.; Eidinoff, M. L.; Forster, W. W. *J. Am. Chem. Soc.* **1939**, *61*, 1539–1543.
- (161) Aston, J. G.; Sagenkahn, M. L.; Szasz, G. J.; Moessen, G. W.; Zuhr, H. F. *J. Am. Chem. Soc.* **1944**, *66*, 1171–1177.
- (162) Majer, V.; Svoboda, V. *Enthalpies of Vaporization of Organic Compounds: A Critical Review and Data Compilation*; Blackwell Scientific Publications: Oxford, U. K., 1985.
- (163) Hales, J. L.; Gundry, H. A.; Ellender, J. H. *J. Chem. Thermodyn.* **1983**, *15*, 211–215.
- (164) Beaton, C. F.; Hewitt, G. F.; Liley, P. E. *Physical Property Data for the Design Engineer*; Hemisphere Pub. Corp.: New York, 1989.
- (165) Goodwin, R. D. *Hydrogen Sulfide Provisional Thermophysical Properties From 188 to 700 K at Pressures to 75 MPa*; Report, NBSIR-83-1694; NTIS No. PB84-122704; 1983
- (166) Russell, H., Jr.; Osborne, D. W.; Yost, D. M. *J. Am. Chem. Soc.* **1942**, *64*, 165.
- (167) Berthoud, A.; Brun, R. *J. Chim. Phys. PCB* **1924**, *21*, 143–160.
- (168) Haines, W. E.; Helm, R. V.; Bailey, C. W.; Ball, J. S. *J. Phys. Chem.* **1954**, *58*, 270–278.
- (169) Haines, W. E.; Helm, R. V.; Cook, G. L.; Ball, J. S. *J. Phys. Chem.* **1956**, *60*, 549–555.
- (170) *Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds American Petroleum Institute Research Project 44*; Carnegie Press, Pittsburgh, PA, 1953.
- (171) *Physical Constants of Hydrocarbons*; ASTM Technical Publication No. 109A, American Society for Testing and Materials: Philadelphia, PA, 1963.
- (172) Yaws, C. L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds (online book)*; Knovel: Norwich, NY, 2003.
- (173) Somsen, G.; Coops, J. *Recl. Trav. Chim. Pays-B.* **1965**, *84*, 985–1002.
- (174) Covington, A. K.; Dickinson, T. *Physical Chemistry of Organic Solvent Systems*; Plenum Press: London, 1973.
- (175) *DMF Product Bulletin*; E. I. duPont, Inc: Wilmington, DE, 1971.
- (176) Gopal, R.; Rigzi, S. A. *J. Indian Chem. Soc.* **1966**, *43*, 179.
- (177) Geller, B. E. *Zh. Fiz. Khim.* **1961**, *35*, 2210.
- (178) Zegers, H. C.; Somsen, G. *J. Chem. Thermodyn.* **1984**, *16*, 225–235.
- (179) Lemire, R. J.; Sears, P. G. *Top. Curr. Chem.* **1978**, *74*, 45–91.
- (180) Wohlfarth, C. *Static Dielectric Constants of Pure Liquids and Binary Liquid Mixtures*; Springer-Verlag: Berlin, 1991; Vol. 6.
- (181) Speight, J. G. *Perry's Standard Tables and Formulas for Chemical Engineers*; McGraw-Hill: New York, 2002.
- (182) Schlundt, H. Ph.D. Thesis, Dept. of Chemistry, University of Wisconsin, Madison, WI, 1901.
- (183) Barthel, J.; Backhuber, K.; Buchner, R.; Hetzenauer, H. *Chem. Phys. Lett.* **1990**, *165*, 369–373.
- (184) Kindt, J. T.; Schmuttenmaer, C. A. *J. Phys. Chem.* **1996**, *100*, 10373–10379.
- (185) Tofts, P. S.; Lloyd, D.; Clark, C. A.; Barker, G. J.; Parker, G. J. M.; McConville, P.; Baldock, C.; Pope, J. M. *Magn. Reson. Med.* **2000**, *43*, 368–374.
- (186) Holz, M.; Mao, X. A.; Seiferling, D.; Sacco, A. *J. Chem. Phys.* **1996**, *104*, 669–679.
- (187) Williams, W. D.; Ellard, J. A.; Dawson, L. R. *J. Am. Chem. Soc.* **1957**, *79*, 4652–4654.
- (188) O'Reilly, D. E.; Peterson, E. M.; Scheie, C. E. *J. Chem. Phys.* **1973**, *58*, 4072–4075.
- (189) Chen, L. P.; Gross, T.; Ludemann, H. D. *Phys. Chem. Chem. Phys.* **1999**, *1*, 3503–3508.
- (190) Hurler, R. L.; Woolf, L. A. *Aust. J. Chem.* **1980**, *33*, 1947–1952.
- (191) Kamei, Y.; Oishi, Y. *B. Chem. Soc. Jpn.* **1972**, *45*, 2437–2439.
- (192) Stevens, E. D. *Acta Crystallogr., Sect. B* **1978**, *34*, 544–551.
- (193) Jeffrey, G. A.; Ruble, J. R.; McMullan, R. K.; Defrees, D. J.; Binkley, J. S.; Pople, J. A. *Acta Crystallogr., Sect. B* **1980**, *36*, 2292–2299.
- (194) Nahringsbauer, I. *Acta Chem. Scand.* **1970**, *24*, 453–462.
- (195) Boese, R.; Blaser, D.; Latz, R.; Baumen, A. *Acta Crystallogr., Sect. C* **1999**, *55*, IUC9900001.
- (196) Craven, B. M.; McMullan, R. K.; Bell, J. D.; Freeman, H. C. *Acta Crystallogr., Sect. B* **1977**, *33*, 2585–2589.
- (197) Golubev, S. N.; Kondrashev, Y. D. *Zh. Strukt. Khim.* **1984**, *25*, 147–150.
- (198) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563–595.
- (199) Wolfenden, R.; Liang, Y.; Matthews, M.; Williams, R. *J. Am. Chem. Soc.* **1987**, *109*, 463–466.
- (200) Wolfenden, R. *Biochemistry* **1978**, *17*, 201–204.
- (201) Abraham, M. H.; Whiting, G. S. *J. Chem. Soc., Perkin Trans. 2* **1990**, 291–300.

CHARMM Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in Polysaccharide and Carbohydrate–Protein Modeling

Olgun Guvench,[†] Sairam S. Mallajosyula,[‡] E. Prabhu Raman,[‡] Elizabeth Hatcher,[‡] Kenno Vanommeslaeghe,[‡] Theresa J. Foster,[†] Francis W. Jamison, II,[†] and Alexander D. MacKerell, Jr.*[†]

[†]Department of Pharmaceutical Sciences, University of New England College of Pharmacy, Portland, Maine 04103, United States

[‡]Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, 20 Penn St., HSF II-629, Baltimore, Maryland 21201, United States

S Supporting Information

ABSTRACT: Monosaccharide derivatives such as xylose, fucose, *N*-acetylglucosamine (GlcNAc), *N*-acetylgalactosamine (GlaNAc), glucuronic acid, iduronic acid, and *N*-acetylneuraminic acid (Neu5Ac) are important components of eukaryotic glycans. The present work details the development of force-field parameters for these monosaccharides and their covalent connections to proteins via O linkages to serine or threonine side chains and via N linkages to asparagine side chains. The force field development protocol was designed to explicitly yield parameters that are compatible with the existing CHARMM additive force field for proteins, nucleic acids, lipids, carbohydrates, and small molecules. Therefore, when combined with previously developed parameters for pyranose and furanose monosaccharides, for glycosidic linkages between monosaccharides, and for proteins, the present set of parameters enables the molecular simulation of a wide variety of biologically important molecules such as complex carbohydrates and glycoproteins. Parametrization included fitting to quantum mechanical (QM) geometries and conformational energies of model compounds, as well as to QM pair interaction energies and distances of model compounds with water. Parameters were validated in the context of crystals of relevant monosaccharides, as well NMR and/or X-ray crystallographic data on larger systems including oligomeric hyaluronan, sialyl Lewis X, O- and N-linked glycopeptides, and a lectin:sucrose complex. As the validated parameters are an extension of the CHARMM all-atom additive biomolecular force field, they further broaden the types of heterogeneous systems accessible with a consistently developed force-field model.

INTRODUCTION

Monosaccharides having the canonical formula $C_n(H_2O)_n$ are essential biomolecular components of life. Examples such as glucose are central to bioenergetics, and their polymers serve both structural and energy-storage functions, with prominent examples including cellulose, starch, and glycogen. However, the role of carbohydrates extends beyond this realm to include biomolecular functions such as molecular recognition. For example, the quality-control mechanism for protein folding,¹ the differences between blood group antigens,² and the ability of viruses to infect host cells^{3,4} all have carbohydrates as critical components. A common theme among the monosaccharides involved in such biomolecular functions is that their atomic compositions differ from the canonical formula. In particular, they are often deoxy, oxidized, or *N*-methylamine derivatives of $C_n(H_2O)_n$ monosaccharides and/or are covalently linked to other biomolecules such as proteins and lipids via bonds involving oxygen or nitrogen atoms.

Classical force field development efforts aimed at enabling accurate modeling of carbohydrates and carbohydrate-containing biomolecular systems have been ongoing for over a decade.^{5–21} While the increased availability of computing resources has allowed for extensive use of quantum mechanical (QM) target data in an effort to capture the conformational energetics of carbohydrates, much of the focus has been on glucose and its diastereomers. Further limiting the scope of force-field-based

carbohydrate modeling is the fact that much of the parameter development work has not been done in the wider context of biomolecular force fields, such that attempts to model heterogeneous biomolecular systems containing proteins, lipids, and/or nucleic acids with carbohydrates may be hampered by differences in force field parametrization protocols and/or functional forms. It is of note that a recent parametrization of hexopyranoses (such as glucose) and their polymers was explicitly made to be compatible with the GROMOS family of biomolecular force fields^{20,22} and also that the most recent iteration of the GLYCAM force field, GLYCAM06,²¹ contains parametrization for carbohydrate derivatives that can form the foundation for a generalizable biomolecular force field.²³

Toward developing a comprehensive additive all-atom carbohydrate force field, we have developed and validated parameter sets for pyranose²⁴ and furanose²⁵ monosaccharides, as well as aldose and ketose linear carbohydrates and their reduced counterparts, the sugar alcohols.²⁶ Parameter sets have also been developed for glycosidic linkages involving both pyranoses^{27,28} and furanoses,²⁸ with the force field shown to reproduce NMR elucidated solution conformational properties of the disaccharides of maltoside and cellobioside.²⁹ Combined, these parameter sets yield a force field that covers most carbohydrates that serve

Received: May 13, 2011

Published: August 09, 2011

Table 1. MD Simulation Details

system	<i>c</i> (Å)	<i>dt</i> (fs)	length (ns)	snapshot frequency (ps ⁻¹)	# of simulations	simulation software
carbohydrate crystals ^a	12	1	4	1	1	CHARMM ^{40,107}
Aqueous Systems						
α-GlcNAc	12	1	10	10	3	CHARMM
β-GlcNAc	12	1	10	10	3	CHARMM
oligomeric hyaluronan	10	2	50	10	5	CHARMM
sialyl Lewis X	10	2	25	10	5	CHARMM
glycoproteins ^{b,c}	12	2	16	5	1	NAMD ⁹⁶
lectin:sucrose ^c	12	2	20	5	1	NAMD ⁹⁶

^a Full listing in Table 2. ^b Full listing in Table S14 (Supporting Information). ^c Langevin thermostating⁴⁶ was used instead of Nosé–Hoover thermostating, and the system was a rectangular prism.

bioenergetic, structural, and energy-storage functions. The present work extends the parameter set to deoxy, oxidized, or *N*-methylamine monosaccharide derivatives as well as covalent linkages to proteins, thereby allowing the simulation of carbohydrates that are important in biomolecular function and molecular recognition. As with the stated previous efforts, the present parameter development was done explicitly in a fashion to make these new models compatible with the CHARMM additive all-atom biomolecular force field for proteins,^{30,31} nucleic acids,^{32,33} lipids,^{34–38} and drug-like small molecules,³⁹ with the intention of creating a widely applicable and robust force field for the modeling of biomolecular systems consisting of any combination of proteins, nucleic acids, lipids, carbohydrates, and/or small molecules.

METHODS

Molecular mechanics (MM) calculations for parameter development were performed with the CHARMM software.^{40,41} The force field potential energy function $U(r)$ was the same as that for the CHARMM protein,^{30,31,42} nucleic acid,^{32,33,43} lipid,^{34–37,44,45} carbohydrate,^{24–28} and small molecule all-atom additive force fields³⁹

$$\begin{aligned}
 U(r) = & \sum_{\text{bonds } b} K_b(b - b_0)^2 + \sum_{\substack{\text{valence} \\ \text{angles } \theta}} K_\theta(\theta - \theta_0)^2 \\
 & + \sum_{\substack{\text{Urey–Bradley} \\ \text{angles } S}} K_S(S - S_0)^2 + \sum_{\text{dihedrals } \chi} K_\chi(1 + \cos(n\chi - \delta)) \\
 & + \sum_{\text{improvers } \varphi} K_\varphi(\varphi - \varphi_0)^2 + \sum_{\substack{\text{nonbonded} \\ \text{pairs } ij}} \varepsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^6 \right] \\
 & + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (1)
 \end{aligned}$$

The first five sums in eq 1 account for bonded interactions. In these sums, K_b , K_θ , K_S , K_χ , and K_φ are bond, valence angle, Urey–Bradley angle, dihedral angle, and improper dihedral angle force constant parameters, respectively. b , θ , S , χ , and φ are the bond distance, valence angle, Urey–Bradley angle 1,3-distance, dihedral angle, and improper dihedral angle values. The subscript 0 indicates an equilibrium value parameter. Additionally, for the dihedral term, n is the multiplicity and δ is the phase angle as in a cosine series. The sum over nonbonded pairs ij

includes a Lennard-Jones (LJ) 6–12 term to account for dispersion and Pauli exclusion and a Coulomb term to account for electrostatic interactions. ε_{ij} is the LJ well depth, $R_{\min,ij}$ is the interatomic distance at the LJ energy minimum, q_i and q_j are the partial atomic charges, and r_{ij} is the distance between atoms i and j . The Lorentz–Berthelot combining rules are used to determine LJ parameters between different atom types.⁴⁶ There is no separate term for hydrogen bonding interactions, as these are accounted for in the parametrization through a combination of LJ and Coulomb energies

A modified version of the rigid three-site TIP3P model was used to represent water,^{47,48} and the SHAKE algorithm⁴⁹ was applied to keep water molecules rigid and to constrain covalent bonds between hydrogens and their covalently bound heavy atoms to their equilibrium values. Gas-phase molecular mechanics energies were calculated using infinite nonbonded cutoffs. Aqueous and crystal simulations employed periodic boundary conditions⁴⁶ to minimize boundary artifacts and to simulate the infinite crystal environment. A force-switched (aqueous) or energy-switched (crystal) smoothing function⁵⁰ was applied to LJ interactions in the range of $c - 2$ to c , where c is the cutoff distance in Ångströms. Long-range Coulomb interactions were handled using particle mesh Ewald⁵¹ with a real-space cutoff of c . The equations of motion were integrated with the “leapfrog” integrator⁵² and a time step of dt . In the molecular dynamics (MD) simulations, the isothermal–isobaric ensemble was generated via Nosé–Hoover thermostating,^{53,54} Langevin piston barostating,⁵⁵ and a long-range correction to the pressure to account for LJ interactions beyond the cutoff distance c .⁴⁶ Condensed-phase simulations were done at experimental temperature and pressure, which was 298 K and 1 atm for all simulations. Simulations of crystals were based on the availability of relevant systems in the Cambridge Structural Database⁵⁶ (CSD) and employed the appropriate experimental unit cell geometries with crystallographic water molecules and/or crystallographic counterions. Aqueous simulations employed a truncated octahedron as the periodic system. For aqueous simulations, the cell length dimensions were varied isotropically to maintain the target pressure during simulation, whereas unit cell edge lengths in crystal simulations were allowed to vary independently. Angular crystal cell parameters of 90° were constrained to this value, while those not 90° were allowed to vary independently. Table 1 lists the simulation c and dt values, along with simulation lengths, MD snapshot frequency, and the number of times each system was simulated. In cases where a system was simulated once, error estimates for data were

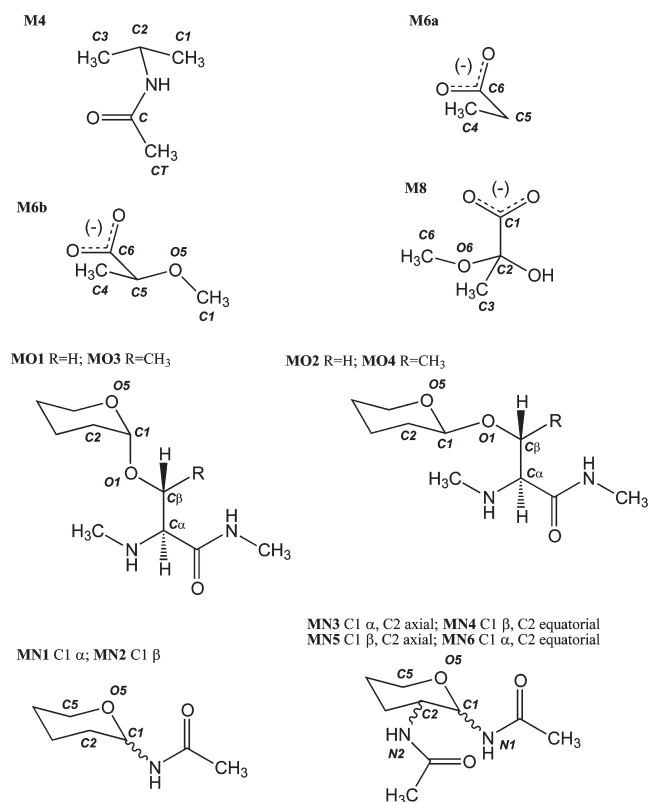


Figure 1. Model compounds used to develop parameters for glucopyranose derivatives GlcNAc and GalNAc (**M4**), glucuronate and iduronate (**M6a** and **M6b**), and sialic acid (**M8**); to develop O-glycosidic linkage parameters involving Ser (**MO1** and **MO2**) and Thr (**MO3** and **MO4**) amino acid side chains; and to parametrize N-linked glycosylation (**MN1**–**6**). Atom labels used in the text are in italics.

generated by treating each MD snapshot as an independent sample and using the expression $t_{\text{critical}} \times s / (n^{0.5})$, where n is the number of snapshots, s is the sample standard deviation, and $t_{\text{critical}} = 1.960$, which is the value for a 95% confidence level for a t distribution with infinite degrees of freedom. In cases where a system was simulated more than once, different trajectories were generated by random assignment of the initial velocities, and each trajectory was treated as an independent sample to generate error estimates for data using the expression $t_{\text{critical}} \times s / (n^{0.5})$, where n is the number of trajectories, s is the standard deviation of the average values calculated for each trajectory, and t_{critical} is the value for a 95% confidence level for a t distribution with $n - 1$ degrees of freedom. Additional simulation details such as system construction are mentioned in the Results and Discussion section for each system.

The Gaussian 03 program⁵⁷ was used for all QM calculations. For small model compounds (Figure 1; **M4**, **M6a**, **M6b**, and **M8**), geometry optimization was done at the MP2/cc-pVTZ level;^{58,59} otherwise, geometry optimization was done at the MP2/6-31G(d) level⁶⁰ followed by a MP2/cc-pVTZ single point energy calculation (MP2/cc-pVTZ//MP2/6-31G(d)). The cc-pVTZ basis set was used for evaluating all conformational energies, as it demonstrates a favorable combination of efficiency and accuracy on carbohydrate systems.^{24,61,62} Vibrational calculations were performed using the MP2/6-31G(d) model chemistry, with tight convergence tolerances applied; the geometries from these unconstrained optimizations were also used as

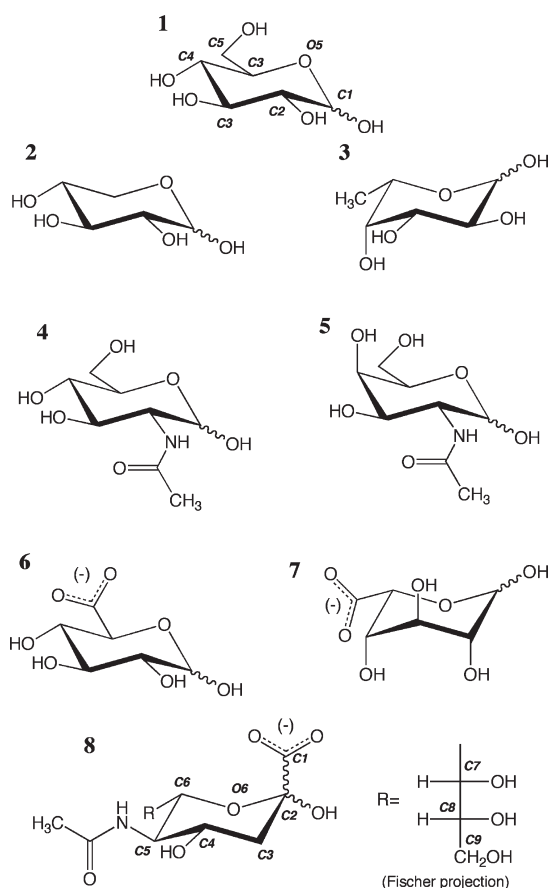


Figure 2. Chemical structures of the pyranose form of glucose; D-glucopyranose (**1**); and related molecules D-xylose (**2**), L-fucose (**3**), N-acetyl-D-glucosamine (GlcNAc) (**4**), N-acetyl-D-galactosamine (GalNAc) (**5**), D-glucuronate (**6**), L-iduronate (**7**), and N-acetyl-D-neuraminic acid (sialic acid) (**8**). Carbon atoms and the ring ether oxygen are labeled in glucose **1**; atom numbering is analogous for **2**–**7**. Sialic acid **8** carbon numbering begins at the carboxyl group. Hydroxyl and carboxyl oxygen atoms derive their numbering from the carbon atom to which they are attached.

reference geometries for gas-phase minimized MM model compound geometries. A scale factor of 0.9434 was applied to QM vibrational frequencies, as required to account for limitations in the level of theory and reproduce experimental frequencies.⁶³ Potential energy decomposition analysis was performed using the MOLVIB utility in CHARMM using the internal coordinate convention of Pulay et al.⁶⁴ All potential energy scans were performed with only the scanned dihedral angles constrained. MM energies were fit to QM potential energies scans using the freely available Monte Carlo simulated annealing (MCSA) dihedral parameter fitting program “fit_dihedral.py”^{27,65} (available for download at <http://mackerell.umaryland.edu>). For each dihedral being fit, three multiplicities n of 1, 2, and 3 were included, and the corresponding K_{χ} values (eq 1) were optimized to minimize the root-mean-square error RMSE between the MM and QM energies as defined by

$$\text{RMSE} = \sqrt{\frac{\sum_i w_i (E_i^{\text{QM}} - E_i^{\text{MM}} + c)^2}{\sum_i w_i}} \quad (2)$$

Table 2. Crystalline Unit Cell Geometries and Volumes^a

CSD code/molecule name	A (Å)			B (Å)			C (Å)			β (deg) ^b			volume (Å ³)		
	exptl	MD	% error ^c	exptl	MD	% error	exptl	MD	% error	exptl	MD	% error	exptl	MD	% error
truncated derivatives															
XYLOSE/xylose	9.25	9.25	0.0	12.67	13.29	4.9	5.64	5.55	-1.6	90	90		660.0	681.0	3.2
ALFUOCO/fucose	14.48	14.63	1.1	7.60	7.95	4.6	6.68	6.43	-3.6	90	90		733.8	747.6	1.9
RHAMAH01/rhamnose	7.91	7.97	0.8	7.92	8.16	3.0	6.67	6.76	1.3	95.6	96.2	0.6	415.9	436.5	5.0
N-acetylaminos															
ACGLUA11/GlcNAc	11.57	11.75	1.5	4.85	4.84	-0.2	9.74	10.05	3.2	116.7	117.3	0.5	488.2	507.3	3.9
AGALAM10/GalNAc	9.16	9.79	6.9	6.32	5.84	-7.6	9.21	9.79	6.4	107.9	109.1	1.2	507.3	527.9	4.1
NACMAN10/ManNAc	7.56	7.65	1.2	7.73	7.76	0.3	18.61	19.10	2.6	90	90		1088.1	1133.2	4.1
acids															
NABDGC/glucuronate	9.21	9.80	6.5	7.01	7.02	0.2	7.38	7.06	-4.3	96.8	97.7	1.4	472.5	479.3	1.4
CANAGL10/galacturonate	13.50	13.19	-2.3	13.50	13.19	-2.3	9.66	9.88	2.3	90	90		1523.7	1489.5	-2.2
KEMYAC/Neu5Ac	7.50	7.50	0.0	7.50	7.50	0.0	29.36	29.08	-1.0	90	90		1652.1	1635.0	-1.0
O-glycans ^d															
COSHEX	7.73	7.79	0.7	8.63	8.61	-0.2	9.94	10.13	1.9	112.5	111.0	-1.3	612.8	634.0	3.5
N-glycans ^d															
AVUVES	6.63	6.90	4.0	19.49	20.63	5.9	8.46	8.02	-5.2	90	90		1093.6	1141.5	4.4
AVUVIW	7.47	7.53	0.8	8.70	8.90	2.2	14.98	15.35	2.5	90	90		973.6	1028.0	5.6
AVUVOC	6.64	6.73	1.4	8.62	8.56	-0.7	15.81	16.28	3.0	90	90		904.3	936.8	3.6
RESJEE	7.86	8.05	2.4	9.42	9.43	0.1	14.01	14.01	0.0	90	90		1038.1	1063.1	2.4
CAKFAV	4.94	4.93	-0.1	7.88	7.94	0.8	17.67	18.50	4.7	91.4	73.5	-19.6	687.8	694.6	1.0
ASGPRS	4.93	4.86	-1.5	24.22	24.46	1.0	7.79	7.76	-0.3	97.7	100.4	2.7	921.8	906.1	-1.7
BEHPIN	4.94	4.92	-0.3	24.26	23.88	-1.6	7.77	7.76	-0.2	97.7	98.1	0.4	922.8	901.9	-2.3
BEHPOT	4.94	4.82	-2.4	16.68	16.31	-2.2	8.08	8.59	6.3	96.1	97.3	1.2	662.0	670.0	1.2

^a MD values are 4 ns averages. The 95% confidence intervals for A, B, and C are <0.02 Å; for β are <0.2°; and for volumes are <0.5 Å³. ^b Constrained to 90° in the simulation if equal to 90° in the experimental crystal, otherwise allowed to vary independently during the simulation. ^c (MD - exptl)/exptl × 100%. ^d Please refer to text for full compound names.

where the sum is over all conformations i of the molecule in the scan, w_i is a weight factor for conformation i , E_i^{QM} is the QM energy of conformation i , E_i^{MM} is the total MM energy, including the energy of the dihedrals for which the parameters are being optimized (eq 1), and c is a constant that vertically aligns the data as the optimization proceeds to minimize the RMSE and is defined by

$$\frac{\partial \text{RMSE}}{\partial c} = 0 \quad (3)$$

w_i values can be empirically chosen to, for example, favor more accurate fitting of low-energy conformations while sacrificing the fit of high-energy ones. In fitting the dihedral parameters, K_χ values were constrained to be no more than 3 kcal/mol, and phase angles δ were limited to 0 and 180° to maintain symmetry of the dihedral potentials about $\chi = 0^\circ$ and thereby ensure applicability of dihedral parameters to both enantiomers of a chiral compound.

To test the nonbonded parametrization for charged or hydrogen bond-forming moieties in model compounds, QM calculations were done to determine interaction energies for model compound:water-molecule pairs. To ensure consistency across the CHARMM additive force field, these calculations followed a standard procedure.⁶⁶ First, the solute:water interaction distance was optimized at the HF/6-31G(d) level, with constraints on all

other degrees of freedom. Here, the water intramolecular geometry in both QM and MM calculations of pair interaction data was that of the TIP3P water model,⁴⁷ and the model compound geometry was one that was previously gas-phase optimized in the QM or CHARMM representation, respectively. Second, following optimization, HF/6-31G(d) interaction energy target data were calculated as $s(E_{\text{pair}} - E_{\text{solute}} - E_{\text{water}})$, with no basis-set superposition-error correction and the empirical scaling factor of s introduced to yield parameters appropriate for a condensed phase force field,^{30,67} where $s = 1.00$ in the case of solutes having moieties with nonzero formal charge and 1.16 otherwise. The interaction distance target data were calculated as the QM-optimized distance minus 0.2 Å, again to yield parameters appropriate for a condensed phase force field.

RESULTS AND DISCUSSION

I. Parameter Development. Truncated Derivatives. The monosaccharides xylose and fucose can be viewed as truncated derivatives of C₆H₁₂O₆ hexopyranoses like glucose (Figure 2, compounds 2, 3, and 1, respectively). Relative to these hexopyranoses, xylose is missing the entire hydroxymethyl group, while fucose, a deoxyhexopyranose, lacks the hydroxyl on the hydroxymethyl group. From a parametrization standpoint, this suggested the immediate transfer of existing parameters. In particular, existing hexopyranose parameters,²⁴ combined with

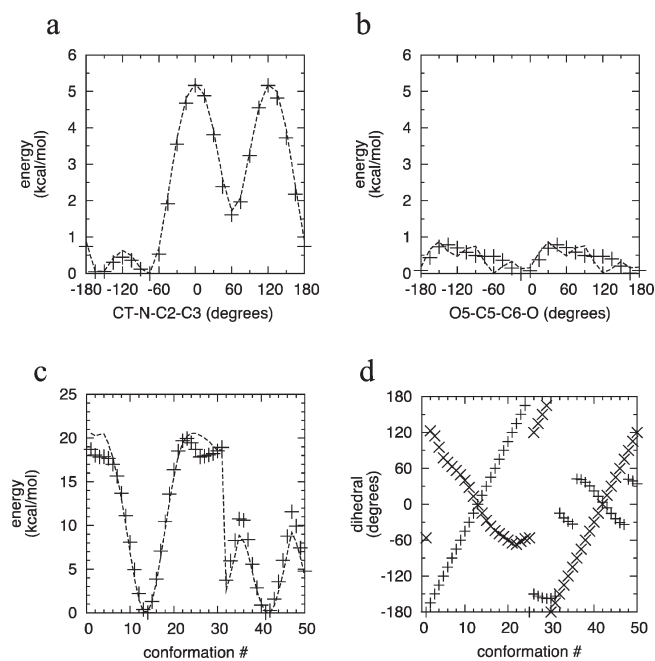


Figure 3. Relaxed potential energy scans of (a) the isopropyl group in compound **M4** and (b) the carboxyl group in compound **M6b**, (c) conformational energies, and (d) hydroxyl and carboxyl dihedral values for compound **M8**. (a–c) MP2/cc-pVTZ energies are represented as crosses and MM energies as a dashed line; (d) hydroxyl dihedral values are represented as crosses and carboxyl dihedral values as \times 's.

existing alkane^{34–38,68} and linear ether parameters⁶⁹ applied to the methyl group on fucose, provided coverage for all atoms and connectivities in these molecules. Crystal simulations of xylose, fucose, and rhamnose (a diastereomer of fucose, with inverted chiralities at C2 and C4) demonstrated the suitability of these transferred parameters. With the exception of the C5–O5 bond length in fucose, all average bond, angle, and dihedral values were consistent with the experimental crystallographic values (Table S1 of the Supporting Information). Further optimization of the C5–O5 bond parameters, which were originally developed for use in hexopyranoses²⁴ and subsequently used in other hexopyranose derivatives (see below), would have required the creation of a new atom type. However, in the interest of balancing accuracy with simplicity and generality of the force field parameters, a new atom type was not introduced, as the extent of disagreement with the crystal data was deemed acceptable. Unit cell geometries were consistent with the experimental values, and in line with prior results,^{24–27} unit cell volumes were systematically overestimated by several percent (Table 2).

N-Acetylaminines. One of the most common modifications to hexopyranoses is the replacement of the C₂ hydroxyl with an *N*-acetylamine group, resulting in monosaccharides like *N*-acetylglucosamine (GlcNAc; Figure 2, 4) and *N*-acetylgalactosamine (GalNAc; Figure 2, 5). In eukaryotes, both GlcNAc and GalNAc are important components of the oligosaccharides that are post-translationally attached to proteins to create glycoproteins.⁷⁰ Using a fragment-based approach, isopropylacetamide (Figure 1, **M4**) was used to develop parameters for these types of sugars. Initial parameters for **M4** were transferred from existing parameters previously developed for *N*-methylacetamide in the context of proteins,³⁰ and with carbon Lennard-Jones parameters taken

Table 3. Solute:Water Pair Interaction Energies and Distances for Model Compounds **M4**, **M6b**, and **M8**

water orientation ^a	energy (kcal/mol)			distance (Å)		
	QM ^b	MM	MM-QM	QM-0.20 ^b	MM	MM-QM
M4						
<i>a</i>	−5.39	−5.86	−0.48	1.97	1.95	−0.02
<i>b</i>	−5.92	−6.05	−0.13	1.94	1.94	0.00
<i>c</i>	−6.74	−6.84	−0.10	1.81	1.76	−0.05
<i>d</i>	−8.44	−7.23	1.21	1.78	1.77	−0.01
average			0.13			−0.02
standard deviation			0.74			0.02
M6b						
<i>a</i>	−14.67	−15.77	−1.09	1.61	1.64	0.03
<i>b</i>	−15.18	−15.01	0.17	2.27	2.22	−0.05
<i>c</i>	−6.08	−2.46	3.61	2.19	2.39	0.19
<i>d</i>	−15.89	−17.64	−1.75	1.85	1.72	−0.13
<i>e</i>	−10.48	−11.05	−0.57	1.77	1.76	−0.01
<i>f</i>	−7.40	−6.38	1.02	1.87	1.84	−0.03
average			0.23			0.00
standard deviation			1.92			0.10
M8						
<i>a</i>	−12.86	−13.23	−0.37	1.65	1.69	0.03
<i>b</i>	−13.47	−13.56	−0.10	2.30	2.23	−0.08
<i>c</i>	−5.27	−2.93	2.34	1.97	2.01	0.04
<i>e</i>	−8.92	−9.10	−0.18	1.86	1.85	−0.02
<i>f</i>	−6.59	−5.70	0.90	1.97	1.93	−0.04
<i>g</i>	−9.78	−9.56	0.22	1.73	1.82	0.10
<i>h</i>	0.08	−0.70	−0.78	2.10	2.29	0.19
<i>i</i>	0.23	−0.74	−0.97	2.16	2.28	0.12
average			0.13			0.04
standard deviation			1.06			0.09

^a Molecular geometries are as illustrated in Figure S2, S4, and S6 of the Supporting Information. ^b HF/6-31G(d) target energies have been scaled by 1.16 for the neutral compounds (**M4** and **M6b**) but not for the charged compound (**M8**), and distances have been shortened by 0.20 Å.

from an improved set of alkane parameters.^{68,69} The 3-fold dihedral term for rotation of the isopropyl group was fit to the QM relaxed potential energy scan (PES), and the transferred C–CT equilibrium length was increased by 0.030 Å. These bonded parameters yielded near-ideal agreement between the QM and the MM conformational energies (Figure 3a), as well as good agreement with bond lengths, valence angles, and dihedral angles (Table S2, Supporting Information). Vibrational frequencies also showed good agreement, with the exception of those that, in the QM representation, involved wagging or deformation of atoms in the amide bond (Figure S1, frequency numbers 10–13, Supporting Information). While these frequencies were overestimated in the MM representation compared to the gas-phase QM, this is in fact appropriate behavior for a condensed-phase force field as the relevant frequencies tend to increase on going from the gas phase to an aqueous environment.³⁰ Finally, using the transferred Lennard-Jones and partial charge non-bonded parameters, water pair interaction energies with the amide CO and NH groups faithfully reproduced the target data (Table 3; Figure S2, Supporting Information).

Transfer of the parameters allowed for immediate creation of models for GlcNAc and GalNAc. Only crystals of the α -anomers of these two sugars were available through the Cambridge Structural Database;⁵⁶ MD simulations of infinite crystals using the lowest *R*-value structures as starting conformations (ACGLUA11, two molecules of α -GlcNAc in unit cell; AGALAM10, two molecules of α -GalNAc in unit cell) showed that the bonds, valence angles, and dihedral angles were all well-represented by the force field model in that average values from the simulations corresponded to those in the reference experimental crystals (Table S3, Supporting Information). The availability of a crystal structure of *N*-acetyl- β -mannosamine (ManNAc) monohydrate (NACMAN10: four molecules of β -ManNAc + four water molecules in unit cell) allowed the testing of a β -anomer using the same parameters and gave similarly good results (Table S3, Supporting Information), and as with the truncated derivatives, crystal volumes for all of the *N*-acetylaminines were overestimated by several percent (Table 2).

Carboxylates. Glucuronate and Iduronate. Oxidation of the C6 alcohol to form a carboxylic acid yields hexopyranose derivatives such as glucuronic acid and iduronic acid, which ionize at physiological pH to yield glucuronate (Figure 2, 6) and iduronate (Figure 2, 7). Among other functions, these compounds are important as components of glycosaminoglycans⁷⁰ and for metabolic conjugation with drugs by the liver.⁷¹ Continuing with a model compound-based approach, **M6a** and **M6b** (Figure 1) were used to develop parameters for glucuronate and iduronate. As detailed below, bonded and nonbonded parameters were transferred from analogous parameters, and missing angle and dihedral parameters were fit to QM geometries and conformational energies.

M6a (propanoate) parameters were previously developed for the CHARMM protein force field,³⁰ allowing for immediate extension to **M6b** (α -methoxy-propanoate), which mimics the C6 carboxylate in the context of the C4, C5, O5, and C1 atoms of the hexopyranose ring. Lennard-Jones parameters were updated on the basis of recent work on alkanes.^{68,69} Additional parameters required upon introduction of the methoxy group were transferred by analogy, leaving only the O5–C5–C6–O dihedrals to be fit. **M6b** was constructed with the C1–O5–C5–C6 analogous dihedral in the *trans* conformation. An optimized MP2/cc-pVTZ scan was done on the OCCO torsion with no other constraints on the system, and the C1–O5–C5–C6 dihedral stayed *trans* for the entire scan. Self-consistent optimization of the O61–C6–O62 angle (equilibrium angle increased by 8°), the C5–C6–O61 and –O62 angles (equilibrium angle decreased by 4°), the O5–C5–C6 angle (equilibrium angle decreased by 8.5° relative to linear ethers⁶⁹), and the OCCO torsions yielded good conformational energies (Figure 3b), minimum-energy geometries (Table S4, Supporting Information), vibrational frequencies (Figure S3, Supporting Information), and water interaction energies (Table 3 and Figure S4, Supporting Information) as compared to the QM target data.

Relevant crystals included β -glucuronate (NABDGC) and α -galacturonate (CANAGLC10). In the case of glucuronate, the full monoclinic unit cell consisted of two monosaccharides, two water molecules, and two sodium ions, and in the case of galacturonate, the full hexagonal unit cell consisted of six monosaccharides, 12 water molecules, two sodium ions, and two calcium ions. After initial simulations, the C5–C6 equilibrium bond length was reduced by 0.042 Å, after which the bonds, valence angles, and dihedral angles had average values from MD

simulations that were consistent with the experimental geometries (Table S5, Supporting Information). An interesting exception was bonds and angles involving oxygens in the carboxyl groups. In the crystals, there is a 2–3% difference between equivalent C6–O bonds and between equivalent C5–C6–O angles, presumably due to differences in the local chemical environments of the carboxyl oxygens. In the MM representation, there is no difference in the parameters for these equivalent oxygens, and these large asymmetries in bonds and angles are not reproduced, pointing to some limitations of the pairwise-additive functional form of the force field where the bonds and angles are treated with harmonic terms. Also of note is that the O5–C5–C6 angles in these crystals have accurate values in the simulations relative to the experiments, demonstrating the transferability of the corresponding parameter optimized on **M6b**. Finally, both crystals had average unit cell parameters from the simulations that were largely consistent with the experimental reference values (Table 2).

Sialic Acid. Like glucuronic acid and iduronic acid, *N*-acetylneuraminic acid (Neu5Ac), or “sialic acid” as it is commonly called, has a carboxylic acid moiety that is deprotonated at physiological pH (Figure 2, 8). Neu5Ac is important not only as a common component of glycosyl groups added as post-translational modifications to proteins but also as a critical participant in molecular recognition resulting in the viral infection of human cells, in particular influenza virus infection.^{4,72} In addition to the carboxyl group, Neu5Ac contains an *N*-acetylamine group, like GlcNAc and GalNAc, as well as a linear polyalcohol group, like linear carbohydrates and sugar alcohols.²⁶ While parameters for the *N*-acetylamine group can readily be transferred from those developed for GlcNAc and GalNAc, and parameters for the linear polyalcohol group from previously developed linear polyalcohol parameters,²⁶ parametrization of the carboxyl group is complicated by the presence of not only an ether moiety connected to the same carbon atom C2 but also a hydroxyl group.

M8 (α -methoxy-lactate), which is **M6b** with the addition of a hydroxyl group, was used to develop parameters involving the C2 anomeric carbon in Neu5Ac. After transferring analogous parameters from **M6b**, both bonded and nonbonded parameters were optimized to reproduce target QM geometries, vibrational frequencies, conformational energies, and water pair interaction energies. In particular, to reproduce angle geometries, equilibrium angle values for the C1–C2–C3, O2–C2–C3, and C3–C2–O6 angles were adjusted, resulting in good agreement with the target data (Table S6, Supporting Information). Angle parameters optimized using **M6b** proved transferable, and as with **M6b**, asymmetries in bond and angle geometries involving the equivalent carboxyl oxygen atoms seen in the QM representation were not captured in the MM representation. The MM vibrational frequencies were consistent with those from QM calculations, with the exception of the hydroxyl OH stretch, which has the highest frequency in the empirical model (Figure S5, Supporting Information). In the force field, this OH stretch frequency is $\sim 700\text{ cm}^{-1}$ greater than the next-highest set of frequencies, which are due to methyl CH stretches, and is similar to calculated QM or experimental infrared OH stretching frequencies for simple alcohols.²⁴ In contrast, in the QM representation, the OH stretch has a frequency similar to the CH stretches and therefore much lower than OH stretching frequencies for simple alcohols. The reason for this is quite clear from the minimum-energy geometry used to compute the

vibrational frequencies, in which the hydroxyl group is oriented to form an intramolecular hydrogen bond with the carboxyl group. This strong hydrogen bond leads to weakening of the OH bond as evidenced by an increase of 0.04 Å in the OH bond length for conformations having this intramolecular interaction relative to conformations without the interaction, as observed in a relaxed QM scan of hydroxyl rotation (described below). This phenomenon cannot be captured using molecular mechanics; however, since covalent bonds involving hydrogens are typically constrained to their equilibrium values using SHAKE⁴⁹ or a related algorithm, this limitation is not a significant concern for intended applications of the model.

QM target conformational energies for **M8** were from geometry-optimized MP2/cc-pVTZ scans of carboxyl and hydroxyl rotation in increments of 15°. The only constrained degree of freedom during the scans was the dihedral being scanned. Therefore, during dihedral scanning of the carboxyl group, the hydroxyl group underwent rotation due to nonbonded interactions with the carboxyl group, and vice versa. In contrast, the C1–C2–O6–C6 dihedral was built in the *trans* geometry and stayed in this local minimum throughout the dihedral scanning. Dihedral parameters for O_{carboxyl}–C1–C2–O2 and C1–C2–O2–HO2 were simultaneously fit to the 50 QM conformational energies/geometries,⁶⁵ with harmonic restraining potentials applied to the O_{carboxyl}–C1–C2–O2, C1–C2–O2–HO2, and C1–C2–O6–C6 dihedrals in the MM representation to ensure a match between the MM and QM conformations during the fitting process. The resulting optimized parameters yielded good agreement with the target data (Figure 3c,d). Finally, water interaction energies with the carboxyl group as a hydrogen bond acceptor (analogous to Figure S4, Supporting Information), the ether as a hydrogen bond acceptor (analogous to Figure S4), and the hydroxyl as both an acceptor and a donor (Figure S6, Supporting Information) showed systematically too favorable interaction energies with the carboxyl group using the partial charges transferred from **M6a/b**, which, as described above, were themselves directly transferred from previous work and seen to be suitable in the contexts of **M6a/b**. The partial charges on the carboxyl oxygens were therefore adjusted from –0.76 to –0.60e and the partial charge on the carboxyl carbon decreased from 0.62 to 0.30e. While this may appear to be a large change, it is important to note that the net charge of –1e on the carboxylate group remains unaltered and that this net charge is the main determinant for the strength of electrostatic interactions with this moiety. Prior to the charge redistribution, QM interactions of the carboxyl oxygens with water were too favorable by ~2 kcal/mol. Additionally, interactions with the ether oxygen were too unfavorable by ~2 kcal/mol owing to electrostatic repulsion of water by the adjacent carboxyl carbon. In contrast, with the new partial charge set, good agreement was achieved with the target data (Table 3). Bonded and nonbonded parameter optimization was done self-consistently, and all presented data are from the final parameter set.

To create a force field model for Neu5Ac **8**, parameters from **M8** were combined with those from hexopyranoses,²⁴ polyalcohols,²⁶ and **M4**. MD simulation of the single example of Neu5Ac in the deprotonated form in the Cambridge Structural Database⁵⁶ (KEMYAC; 4 molecules of monosaccharide in the β anomeric form + 4 sodium ions + 12 water molecules, in the complete tetragonal unit cell) pointed to additional parameter optimization. As with the analogous parameter in crystals for β-glucuronic acid and α-galacturonic acid, the C1–C2

equilibrium bond length was reduced by 0.042 Å; additionally, the C2–C3 equilibrium bond length was increased by 0.035 Å, the O2–C2–O6 equilibrium angle value was reduced by 3.5°, and the C3–C2–O6 equilibrium angle value was increased by 2.5° in order to achieve good agreement of bonds, angles, and geometries at and near the carboxylate moiety (Table S7, Supporting Information). Finally, geometric parameters for the rectangular unit cell were in very close agreement with the experimental values (Table 2).

O-Glycans. The O-glycosidic carbohydrate–protein bond, between the anomeric carbon of a carbohydrate and the side chain alcohol of either the amino acid serine (Ser) or threonine (Thr),^{70,73} is an important biological linkage present, for example, in mucin glycoproteins⁷⁴ and as a post-translational modification for cytosolic proteins in signaling pathways.⁷⁵ To develop force field parameters for O-glycosidic linkages, four dipeptide derivatives of the hexopyranose analog tetrahydropyran were chosen as model compounds. These model compounds correspond to the α and β anomers of Ser- (Figure 1, **MO1** and **MO2**) and Thr-linked (Figure 1, **MO3** and **MO4**) carbohydrates. Initial values for bonded and nonbonded parameters were transferred from ethers,⁶⁹ carbohydrates,²⁴ and proteins,³⁰ leaving as targets for parametrization the glycosidic dihedrals about the C1–O1 bond (O5–C1–O1–Cβ and C2–C1–O1–Cβ), O1–Cβ bond (C1–O1–Cβ–Cα and additionally C1–O1–Cβ–Cγ in the Thr-linked analogs) and the Cβ–Cα bond (O1–Cβ–Cα–N and O1–Cβ–Cα–C).

To parametrize glycosidic dihedral rotation about the C1–O1 bond and the O1–Cβ bond, 2D MP2/cc-pVTZ//MP2/6-31G(d) scans were performed on the O5–C1–O1–Cβ (φ_s)/C1–O1–Cβ–Cα (ψ_s) surfaces for all four model compounds (Figure 4). Global minima for the α anomers were located at φ_s/ψ_s = –105°/60° and –90°/45° for the Ser and Thr dipeptides (**MO1** and **MO3**), respectively. For the β anomers, global minima corresponded to φ_s/ψ_s values of –75°/75° and –75°/45° for the Ser and Thr dipeptides (**MO2** and **MO4**), respectively. During the optimized dihedral scans, the only constraints were on the dihedrals being scanned, thereby allowing full relaxation of all other degrees of freedom. For example, the peptide backbone geometry relaxed to various parts of the extended region of the protein φ/ψ Ramachandran surface for each of the four global-minimum structures, with QM-optimized φ/ψ values for the Ser α,β anomers **MO1** and **MO2** being –156°/168° and –157°/161°, respectively, and for the Thr α,β anomers **MO3** and **MO4** being –153°/170° and –153°/172°, respectively.

For the Ser analogs **MO1** and **MO2**, dihedral parameters for O5–C1–O1–Cβ, C2–C1–O1–Cβ (both involving rotation about the same bond C1–O1), and C1–O1–Cβ–Cα were simultaneously fit to both QM potential energy scans. Similarly, for the Thr analogs **MO3** and **MO4**, parameters for the dihedrals O5–C1–O1–Cβ and C2–C1–O1–Cβ (both involving rotation about the same bond C1–O1) and for C1–O1–Cβ–Cα and C1–O1–Cβ–Cγ (both involving rotation about the same bond O1–Cβ), were simultaneously fit to both QM scans.

To parametrize the dihedral rotations O1–Cβ–Cα–N and O1–Cβ–Cα–C about the Cβ–Cα bond, 1D MP2/cc-pVTZ//MP2/6-31G(d) scans were performed on the O1–Cβ–Cα–N dihedral for all four model compounds (Figure S7, Supporting Information). Global minima for the α anomers were located at –60° and 75° for the Ser and Thr dipeptides (**MO1** and **MO3**), respectively. For the β anomers, global minima corresponded to dihedral values of –165° and 45° for the Ser and Thr dipeptides

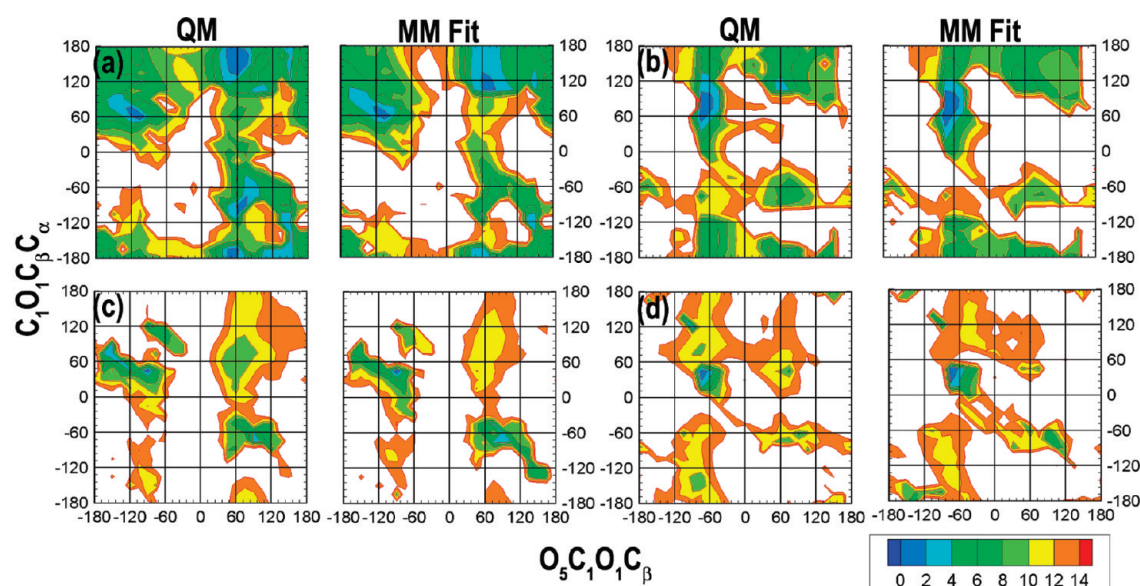


Figure 4. 2D-Dihedral potential energy scans on the $O5-C1-O1-C\beta$ (ϕ_s) and $C1-O1-C\beta-C\alpha$ (ψ_s) dihedrals of the model compounds (a) **MO1**, (b) **MO2**, (c) **MO3**, and (d) **MO4** representative of the O-glycan linkages. Energies are in kcal/mol, with contours every 2 kcal/mol. Only energies below 14 kcal/mol have been plotted for the sake of clarity.

(**MO2** and **MO4**), respectively. For both the Ser and Thr analogs, **MO1** to **MO4**, dihedral parameters for $O1-C\beta-C\alpha-N$ and $O1-C\beta-C\alpha-C$ were simultaneously fit to both the anomeric QM potential energy scans. To ensure faithful reproduction of conformational energies near the QM minima, during the MCSA fitting, the local minimum conformations were given weight factors w_i (eq 2) of 3; conformations with energies above 14 kcal/mol, weight factors of 0; and all other conformations, weights of 1.

A single relevant crystal structure, namely that of a Thr α -anomer (CSD code, R factor, compound name: COSHEX, 3.8, O- α -D-mannopyranosyl-(1-3)-L-threonine), was found through a CSD search, and this structure along with QM-optimized structures of **MO1**–**4** were used to guide additional parametrization of bonded terms. MD simulations of the crystal with two monomers per unit cell showed that a few transferred equilibrium bond lengths and valence angles had to be modified to better match the experimental intramolecular geometries. Therefore, the $O1-C\beta$ equilibrium bond length was increased by 0.01 Å, the equilibrium valence angles for $O1-C\beta-C\alpha$ and $O1-C\beta-C\gamma$ were decreased by 4.5° and 1.5°, and the equilibrium valence angle for $C\beta-O1-C1$ was increased by 2.0°. These additional optimizations yielded good reproduction of both QM and crystal geometries (Table S8, Supporting Information) and crystal unit cell dimensions consistent with other crystals in the present study (Table 2).

Of note, the O-glycan parametrization was also able to reproduce the correct anomeric configurational preference when compared with QM calculations. Defining the anomeric ΔE as $E_\alpha - E_\beta$, for the Ser O-linkage, the MM ΔE value of 1.72 kcal/mol compares reasonably well with the QM value of 3.17 kcal/mol. Similar agreement is seen for the Thr O linkage, where the MM ΔE value is -3.28 kcal/mol and the QM value is -5.27 kcal/mol. Thus, we note that the force field is able to predict the correct configurational preferences for both the Ser and Thr linkages with the Ser linkage favoring the α configuration and the Thr linkage favoring the β configuration. The final potential energy surfaces using the final optimized parameters are

presented in Figure 4a–d and Figure S7 (Supporting Information), with RMSE values with respect to the QM conformational energies ranging from 0.9 to 2.0 kcal/mol across all of the dihedral scans.

N-Glycans. Post-translational protein modification by N-linked glycosylation consists of the addition of oligosaccharides to the side chains of asparagine (Asn) residues. This covalent modification, which occurs in the endoplasmic reticulum, plays a critical role in cell surface expression and is often required for protein stability and biological function.^{70,76} It has been found that N-linked glycosylation generally occurs at the sequence Asn-X-Ser/Thr, where X is any amino acid except proline.^{77,78} This type of glycosylation is found in nearly all eukaryotes,^{70,76} and in most cases the linkage occurs between Asn and N-acetylglucosamine (GlcNAc), replacing the alcohol moiety of GlcNAc C1 with an amide linkage to the Asn side chain.

To develop the force field parameters for this linkage, tetrahydrofuran with N-acetylamino substituted at C1 was chosen as the model compound, with both α and β anomer analogs used for the parametrization process (Figure 1, **MN1** and **MN2**). Most of the initial bond, angle, and dihedral parameters were readily transferred from the N-acetylamino substitution at the C2 position, as developed for GlcNAc and GalNAc. Additionally, the $O5-C1-N$ angle parameter was transferred from $O5-C1-O$ by analogy, and the dihedral parameters $C5-O5-C1-N$ and $O5-C1-N-C$ were transferred from those for $C5-O5-C1-O$ and $C1-C2-N-C$.

To test the transferred parameters QM MP2/cc-pVTZ//MP2/6-31G(d) scans were performed for the $O5-C1-N-C$ dihedral in the two model compounds (Figure 5a,b). These scans were followed by two additional scans of the $C5-O5-C1-N$ dihedral with the $O5-C1-N-C$ dihedral constrained to its corresponding QM global minimum (Figure 5c,d). In the case of the $C5-O5-C1-N$ scans, which correspond to ring deformation, the transferred parameters (Figure 5c,d “MM Trsf’d”) adequately reproduced the target data and could not be further improved by additional fitting (Figure 5c,d “MM Fit”).

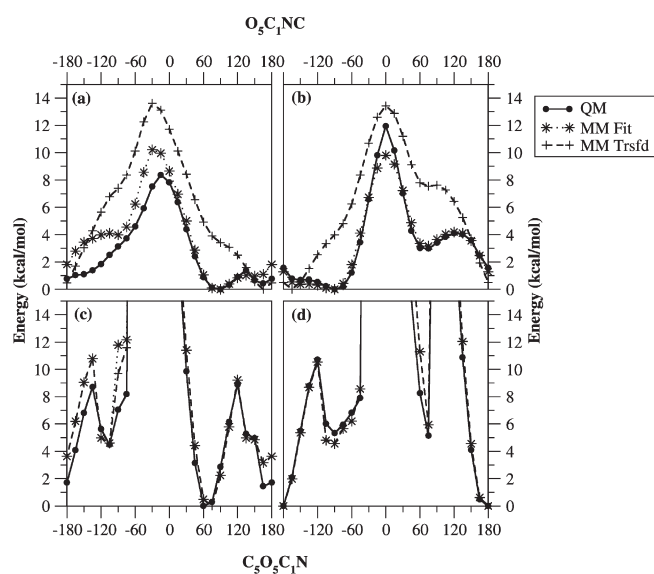


Figure 5. Dihedral potential energy scans for model compounds MN1 and MN2 representative of the *N*-glycan linkages. (a) MN1, O5–C1–N–C; (b) MN2, O5–C1–N–C; (c) MN1, C5–O5–C1–N; (d) MN2, C5–O5–C1–N.

In contrast, the transferred parameters for O5–C1–N–C, which determine the energetics of rotation of the *N*-acetylamino group, gave incorrect locations for the minimum energies as well as barriers to rotation that were too high (Figure 5a,b “MM Trsfd”). Using the QM O5–C1–N–C scans for both model compounds as target data, dihedral force constants were developed for the O5–C1–N–C dihedral using the MCSA fitting procedure, with the same weighting protocol described previously for the O-glycan model compounds. After fitting, RMSE values for the MN1 and MN2 O5–C1–N–C scans were 1.36 and 0.58 kcal/mol, respectively, reflecting a much closer match to the target QM surfaces (Figure 5a,c “MM Fit”) than with the transferred parameters.

To test the transferred and optimized parameters, geometrical descriptors of the QM and MM minimized geometries were compared along with crystalline intramolecular geometries and unit cell parameters for C1 monosubstituted monosaccharides. A CSD survey yielded four monosubstituted *N*-acetylamino crystals (CSD ref code, R factor, compound name: AVUVES, 4.02, β -1-*N*-acetamido-*D*-mannopyranose monohydrate, AVUVIW, 3.82, β -1-*N*-acetamido-*D*-galactopyranose, AVUVOC, 3.86, β -1-*N*-acetamido-*D*-xylopyranose, RESJEE, 3.28, β -1-*N*-acetamido-*D*-glucopyranose). The results of the comparison between the QM and MM intramolecular geometries and the MD simulations of the infinite crystals are tabulated in Tables S9 and S10 of the Supporting Information and demonstrate good agreement of the MM data with regard to both QM-optimized model compounds’ geometries and crystal data for bonds and angles. MM dihedral angles are also consistent with target QM and crystal data with the exception of those involving rotation around the C1–N bond in the crystal. However, for these O5–C1–N–C and C2–C1–N–C dihedrals, discrepancies between the crystallographic and MD average values can be explained as resulting from a flat energy profile in the region of the global minima. In particular, for the β anomer analog MN2, the energy cost for going from a O5–C1–N–C dihedral value of -90° to one of -120° is only 0.5 kcal/mol (Figure 5b). And finally, as with other crystal

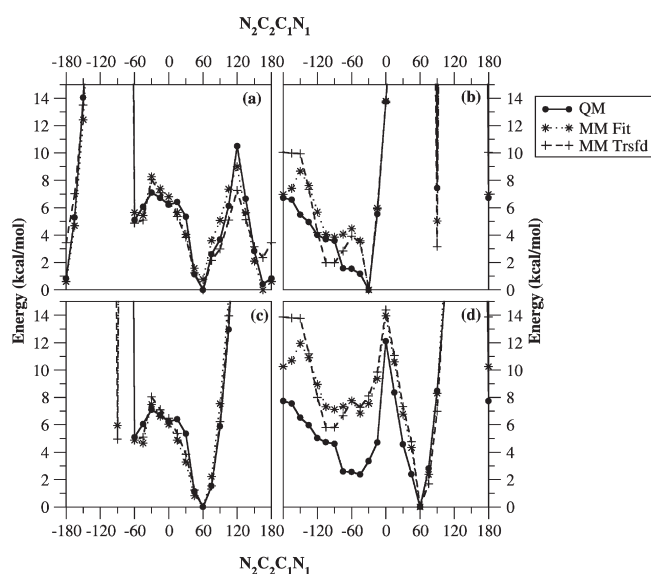


Figure 6. Dihedral potential energy scans about the N2–C2–C1–N1 dihedral for model compounds (a) MN3, (b) MN4, (c) MN5, and (d) MN6.

simulations, a systematic overestimation of the crystal volumes was observed (Table 2).

Since *N*-glycosylation commonly involves linkage of *N*-acetylglucosamine (GlcNAc) to the side chain of Asn, parameters are required for the dihedral angle between the nitrogens of the *N*-acetylamino groups at positions C1 (anomeric carbon) and C2 of GlcNAc involved in such a linkage. To parametrize this dihedral, tetrahydropyran with *N*-acetylamino substitutions at both the C1 and C2 positions was chosen as the model compound, and QM conformational energies were collected for all four possible diastereomers (Figure 1, MN3, MN4, MN5, MN6). Initial parameters for N2–C2–C1–N1 were transferred from the analogous OCCO dihedral of the hexopyranose monosaccharide force field, having been developed using ethylene glycol as a model compound and validated on the basis of crystallographic ring pucker geometries.²⁴ These transferred parameters (Figure 6, “MM Trsfd”) reproduce the QM energy scans (Figure 6, “QM”) as well as the parameters explicitly fit to the target data (Figure 6, “MM Fit”), including good reproduction of the locations and shapes of global energy minima. Reflecting the appropriateness of the transferred dihedral parameters are the similarities in the RMSE values of the transferred vs fit parameters for MN3, MN4, MN5, and MN6: 0.88 kcal/mol vs 0.70 kcal/mol, 1.86 kcal/mol vs 1.73 kcal/mol, 0.43 kcal/mol vs 0.44 kcal/mol, and 3.12 kcal/mol vs 3.41 kcal/mol. Thus, the transferred parameters were retained as the final parameters and were subsequently used to compare the geometries of the QM global minimum and the MM minimized geometries (Table S11, Supporting Information). The MM model, which uses a single set of parameters for all four model compounds, faithfully captures bond lengths and angles, thereby requiring no further adjustment of the equilibrium bond lengths and angles, as well as most dihedrals. The exception is for the O5–C1–N–C and C2–C1–N–C dihedrals, where average errors are 20.9° and 18.0° , in part due to the flat potential profile associated with this dihedral as discussed above. Furthermore, these average errors are heavily influenced by one outlier, namely MN6, because the MM optimized geometry of MN6 favors an intramolecular

hydrogen bond between the two acetamine units, thereby locking the O5–C1–N–C dihedral angle at 153° compared to the QM value of 83°. Excluding this compound from the analysis yields average errors of 4.8° and 0.3° for these two dihedrals. A CSD survey yielded only one disubstituted *N*-acetamine crystal (CSD ref code, R factor, compound name: CAKFAV, 2.50, *N'*-(2-acetamido-2-deoxy- β -D-glucopyranosyl) acetamide monohydrate). On the basis of an MD simulation of the crystal, all bond lengths and valence angles are well reproduced by the transferred force field parameters (Table S12, Supporting Information), with average errors for bond lengths of 0.008 Å and ranging from -1.1° to $+2.1^\circ$ for valence angles. Furthermore, all dihedral angles, including those for rotation about both the C1–N and C2–N bonds, are well reproduced by the transferred parameters. Finally, percentage errors for the unit cell parameters *A*, *B*, *C*, and β were calculated to be -0.1% , 0.8% , 4.7% , and -19.6% , respectively, and the error in the crystal volume was 1.0% (Table 2). The large change in β is due to a slight shift of the two monomers in the crystal with respect to each other; however, this change does not lead to a significant change in the unit cell volume.

To test the applicability of the above parametrization for amino acid–carbohydrate conjugates, the optimized parameters were applied to crystalline *N*-linked monosaccharides. A CSD survey yielded three crystal structures of *N*-linked monosaccharides, all with the *N*-linkage in the β conformation (CSD ref code, R factor, compound name: ASGPRS, 6.00, 2-Acetamido-1-*N*-(*L*-aspart-4-oyl)-2-deoxy- β -D-glucopyranosylamine hydrate, BEHPIN, 5.40, 4-*N*-(2-acetamido-2-deoxy- β -D-glucopyranosyl)-*L*-asparagine trihydrate, BEHPOT, 7.20, 4-*N*-(β -D-glucopyranosyl)-*L*-asparagine monohydrate). In all cases, the MD simulations reproduced crystallographic bond lengths, valence angles, and dihedral angles to within acceptable errors (Table S13, Supporting Information), with the exception of the C1–C2–N2–C and C3–C2–N2–C dihedrals. These latter dihedrals correspond to rotation about the N2–C2 bond, and the observed error can be rationalized from the potential energy surface for the model compound **M4** (Figure 3a), where the conversion from one minimum to another in the crystal simulation corresponds to conversion across the 0.5 kcal/mol barrier from one of the two global minima to the other in **M4**. And finally, all unit cell parameters as calculated by averaging across the MD trajectories were close to the corresponding experimental crystallographic values (Table 2).

II. Application to Example Systems. Toward demonstrating the utility of the new parameter set, MD simulations were done on relevant carbohydrates alone and covalently conjugated or noncovalently bound to proteins. Carbohydrate-only systems consisted of aqueous simulations of monomeric GlcNAc, the linear glycosaminoglycan polymer hyaluronan, and the branched glycan sialyl Lewis X. Covalent carbohydrate–protein conjugates consisted of two glycoproteins containing only *N*-linked glycans and two glycoproteins having both *N*-linked and *O*-linked glycans. Finally, a MD simulation was performed on sucrose noncovalently bound to a lectin.

Conformational Properties of the GlcNAc Acetamido Group. NMR studies on GlcNAc allow for testing the ability of the force field to reproduce conformational sampling of this sugar in solution. On the basis of the potential energy scan for **M4** (Figure 3a), the acetamido group of GlcNAc is anticipated to have three stable conformations. Using previously developed Karplus equations for three-bond *J*-coupling $^3J(\text{H}^{\text{N}}\text{H}^2)$ between the protons on C2 and the amide nitrogen,⁷⁹ it is possible to

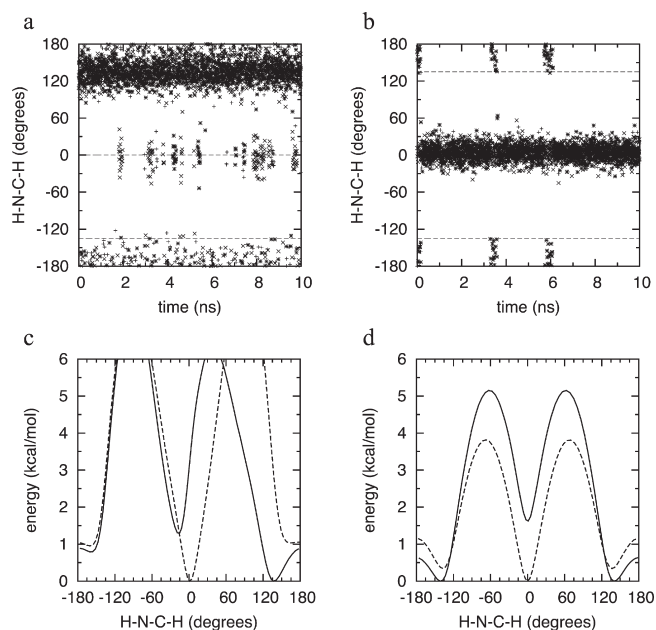


Figure 7. Conformational properties of the GlcNAc acetamido group. $\theta_{\text{H-N-C-H}}$ vs time is shown for the α - (a) and β -anomers (b) in three independent aqueous MD simulation trajectories (+, ×, *) for each anomer, along with dashed lines at $-135^\circ/0^\circ/+135^\circ$. Also shown are (c) the gas-phase potential energy surfaces for $\theta_{\text{H-N-C-H}}$ in α - (solid line) and β -GlcNAc (dashed line) in the absence of electrostatic interactions and (d) the gas-phase potential energy surfaces for $\theta_{\text{H-N-C-H}}$ for **M4** with (solid line) and without (dashed line) electrostatic interactions.

compare the conformational properties of this moiety in aqueous MD simulations with NMR data. The relevant H–N–C2–H dihedral value θ can be related to the **M4** CT–N–C2–C3 dihedral scan data by the relationship $\theta_{\text{H-N-C-H}} = \theta_{\text{C-N-C-C}} - 60^\circ$, which yields three minima for $\theta_{\text{H-N-C-H}}$: one at 0° (*cis*) and two energetically equivalent minima on either side of 180° (*trans*[−] and *trans*⁺), with the *trans* minima 1.7 kcal/mol more stable than the *cis* minimum in the MM representation (Figure 3a). $\theta_{\text{H-N-C-H}}$ values from three 10 ns trajectories each of the α - and β -anomers of GlcNAc show very different behavior for the two anomers. In the case of the α -anomer, the population of sampled states is *trans*[−] \gg *trans*⁺ > *cis*, where overall *trans* > *cis*, as anticipated from the vacuum potential energy surface of the model compound, but with a clear energetic asymmetry introduced between *trans*[−] and *trans*⁺ (Figure 7a). The deviation from the vacuum potential energy surface of **M4** is even more striking in the case of the β -anomer, in which the population of sampled states is *cis* \gg *trans* (Figure 7b); it is worth emphasizing that all force field parameters are exactly the same for the two anomers. Interestingly, in cases where *trans* is undersampled relative to the **M4** surface, namely *trans*⁺ for the α -anomer and both *trans*[−] and *trans*⁺ for the β -anomer, the sampled *trans* conformations deviate from the **M4** ideal values of *trans*[−]/*trans*⁺ = $+135^\circ/-135^\circ$ (dashed lines in Figure 7a, b).

Visualization of the trajectories did not point to any obvious stabilizing or destabilizing interactions as causing the difference between the acetamido *cis/trans* conformational preferences between the two anomers. However, the populations may be rationalized in the context of the rotational profiles of the acetamido group in the absence of electrostatic interactions

(Figure 7c). Without electrostatic interactions, there is no possibility of intramolecular hydrogen bonding or electrostatic repulsion between the acetamido group on C2 and the hydroxyl groups attached to C1 and C3, and to a very rough approximation, this mimics the electrostatic shielding in aqueous solution. The energy of the β -anomer as a function of H–N–C2–H has a global minimum in the *cis* conformation, with a broad local minimum of +1 kcal/mol at the *trans* conformation (Figure 7c). This energy surface explains the preference for the *cis* conformer along with sampling of the ideal *trans* conformation instead of the *trans*⁺ and *trans*[−] conformations (Figure 7b). Likewise, for the α -anomer, the vacuum potential energy surface in the absence of electrostatic interactions mirrors the acetamido conformational sampling for aqueous α -GlcNAc (Figure 7a,c). In particular, the surface is no longer symmetric about H–N–C2–H = 0°, because, unlike β -GlcNAc, which has both C1 and C3 hydroxyls in equatorial configurations, the C1 hydroxyl is axial. As a result, the combination of bonded and LJ force field terms yields a global minimum at *trans*[−], which is the conformational state preferentially sampled by β -GlcNAc. Finally, to highlight the importance of electrostatics, it is worth noting that in the absence of electrostatic interactions, the *cis* H–N–C–H conformation for M4 becomes the global minimum by 0.4 kcal/mol relative to the *trans*[±] local minima, whereas when the full force field representation is used, the *trans*[±] conformations are 1.7 kcal/mol more stable than the *cis* (Figure 7d).

Mobli and Almond recently developed Karplus equations specifically for the α - and β -anomers of GlcNAc,⁷⁹ where $\langle^3J(\text{H}^N\text{H}^2)$ for the α -anomer is described by

$$J = 9.56\cos^2(\theta) - 1.62\cos(\theta) + 0.69 \quad (4)$$

and for the β -anomer by

$$J = 9.45\cos^2(\theta) - 2.08\cos(\theta) + 0.63 \quad (5)$$

Using the above equations to calculate ensemble average coupling values (\langle^3J) from the MD trajectories yields values of 7.0 ± 0.1 and 7.7 ± 0.4 Hz (where errors are 95% confidence intervals using the \langle^3J) from each trajectory as an independent sample) for the α - and β -anomers, respectively. In comparison, the values from NMR experiments are 8.9 and 9.1 Hz.⁷⁹ Equation 4 for the α -anomer has maxima of 11.9 Hz at $\pm 180^\circ$ and 8.6 Hz at 0° and a minimum of 0.7 Hz at $\pm 90^\circ$. Similarly, eq 5 for the β -anomer has maxima of 12.2 Hz at $\pm 180^\circ$ and 8.0 Hz at 0° and minima of 0.6 Hz at $\pm 90^\circ$. Assuming that the above QM-derived Karplus equations are appropriate, interpretation of the MD data is complicated by the fact that both the *cis* and *trans* conformations correspond to maxima in eqs 4 and 5. Thus, the underestimation of the \langle^3J values from the MD simulations relative to NMR may arise from local structural deviations from idealized *cis* or *trans* geometries, from inaccurate sampling of *cis* vs *trans* conformational populations, or from a combination of these factors. Using a similar MD protocol with a different force field, the developers of the above equations computed \langle^3J values of 8.9 and 10.4 Hz for the two anomers.⁷⁹ While their computed value for the α -anomer exactly reproduces their NMR value, the computed β -anomer value is overestimated by nearly the same amount as it is underestimated in the present study. More importantly, in the previous work, the α -anomer exclusively sampled the *trans* conformation, and the β -anomer preferentially sampled the *trans* conformation vs *cis* by a factor of 9 to 1. This is in contrast to the

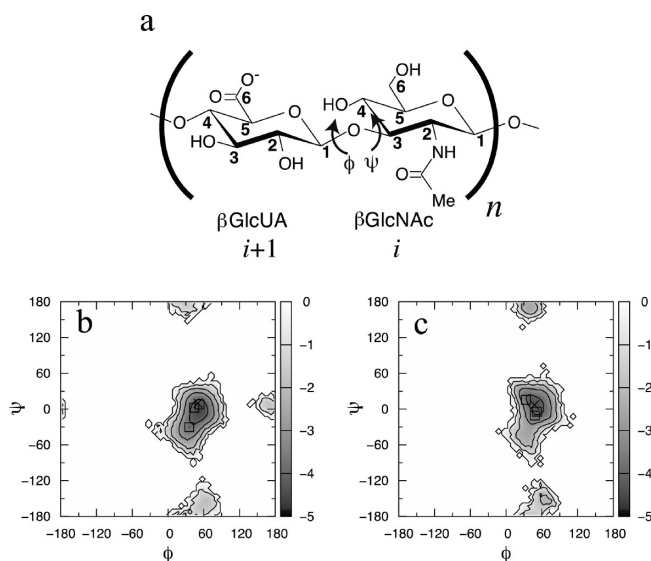


Figure 8. Conformational properties of the glycosidic linkages in hyaluronan. The chemical structure of hyaluronan is shown (a), demonstrating the GlcUA- $\beta(1\rightarrow3)$ -GlcNAc linkage. To either side of this linkage are GlcNAc- $\beta(1\rightarrow4)$ -GlcUA linkages, and monosaccharide residues are numbered starting at the reducing end of the polymer. Boltzmann-inverted hyaluronan 8-mer ϕ/ψ probability distributions for GlcUA- $\beta(1\rightarrow3)$ -GlcNAc (b) and GlcNAc- $\beta(1\rightarrow4)$ -GlcUA glycosidic linkages (c) are shown (in kcal/mol; contour lines are every 1 kcal/mol). $\phi = \text{H1-C1-O}_{\text{link}}-\text{Cx}$ and $\psi = \text{C1-O}_{\text{link}}-\text{Cx}-\text{Hx}$, and data have been aggregated across all linkages of the same type in the 8-mer and across five independent 50 ns simulations. ϕ/ψ angles from NMR⁸¹ and X-ray crystallographic⁸³ structures of hyaluronan are shown as \times 's and open squares, respectively.

present work, where the α -anomer does demonstrate some sampling of the *cis* conformation and where the β -anomer samples the *cis* conformation almost exclusively (Figure 7). With regard to the significant preference of the β -anomer for the *cis* conformation observed here, it is unlikely that it is due to kinetic trapping, as all three trajectories were minimized, heated, and equilibrated with positional restraints on the monosaccharide atoms such that the acetamido group maintained a *trans*⁺ geometry during these initial phases of the simulation. Thus, it remains an open question as to whether or not the β -GlcNAc acetamido group prefers the *cis* or *trans* conformation in aqueous solution, or some combination of the two, as all of these possibilities are consistent with the experimental \langle^3J value.

Conformational Properties of Oligomeric Hyaluronan. The linear glycosaminoglycan hyaluronan, composed of GlcNAc and glucuronate (GlcUA) residues, is an important component of the extracellular matrix and plays structural as well as molecular-recognition roles in biology.^{80,81} The component monosaccharides of hyaluronan are linked together in the repeating motif, ... GlcUA- $\beta(1\rightarrow3)$ -GlcNAc- $\beta(1\rightarrow4)$ -GlcUA... (Figure 8a), and the linear polymer can reach molecular weights of over 1 million Daltons.^{80,81} Recently, the aqueous structure of hyaluronan oligomers has been deduced via NMR spectroscopy as being close to a contracted left-handed 4-fold helix,⁸² and the ϕ/ψ dihedral angles of this repeating structure are maintained when hyaluronan oligomers form complexes with the hyaluronan-binding domain (HABD) of the cell-surface protein CD44.⁸³ Using the hyaluronan 8-mer (HA8; Figure 8a, $n = 4$) coordinates from the A-form HABD:HA8 complex [PDB ID: 2JQC⁸³],

Table 4. Comparison of Hyaluronan Oligomer Acetamido $^3J(\text{H}^{\text{N}}\text{H}^2)$ Values (Hz) from Simulations with NMR Experiments

	experiment ⁸²		MD ^a
	HA4	HA6	HA8
reducing end	9.8	9.7	8.1 ± 0.1 (85% ± 3%)
core ^b	n/a	9.8	9.9 ± 1.3 (17% ± 46%)
	n/a	n/a	10.3 ± 0.1 (1% ± 3%)
nonreducing end	9.7	9.7	10.1 ± 0.7 (10% ± 29%)

^a MD $^3J(\text{H}^{\text{N}}\text{H}^2)$ values are calculated as the average $\langle ^3J \rangle$, where $\langle ^3J \rangle$ is the ensemble-average $^3J(\text{H}^{\text{N}}\text{H}^2)$ in one 50 ns MD simulation, and there are five $\langle ^3J \rangle$'s, one from each simulation. Values in parentheses are the average % of conformations sampled that are *cis*. Error values are 95% confidence intervals calculated as $2.78(\text{average}(\langle ^3J \rangle) \text{ or } \% \text{ cis})/\text{sqrt}(5)$. % *trans* = 1 - % *cis*, and confidence intervals for % *trans* are identical to those for % *cis*. ^b Refers to GlcNAc residues that are neither at the reducing end (residue $i = 1$) nor at the nonreducing end (residue $i = 2n - 1$). In the case of HA8, these are listed in order from the reducing end to the nonreducing end.

five 50 ns simulations of aqueous HA8 were performed. The crystallographically unresolved coordinates of residue GlcUA8 were generated using force field default internal geometries for the monosaccharide, in which the hexopyranose ring is in the energetically favored $^4\text{C}_1$ chair conformation, and GlcUA8- $\beta(1\rightarrow3)$ -GlcNAc7 ϕ/ψ dihedral angles' geometries were constructed in accord with the NMR/crystallographic conformations. The HA8 molecule was centered in a truncated octahedron with sufficient water molecules so that it was at least 10 Å from the nearest edge of the system, overlapping water molecules were deleted, and four sodium ions were added at random positions to achieve a system of net neutral charge. The system was briefly minimized, heated, and equilibrated with harmonic restraints on HA8 atomic positions, and then the system was simulated with a time step dt of 0.002 ps and a cutoff value c of 10 Å with only a harmonic restraining potential on the HA8 center-of-mass, with other MD details per the Methods section. Five independent trajectories were achieved by the random assignment of velocities to the same system at the start of the five separate simulations.

HA8 contains four GlcNAc residues, allowing for the calculation of four separate ensemble-average $^3J(\text{H}^{\text{N}}\text{H}^2)$ values $\langle ^3J \rangle$, one for each residue, using the Karplus relationship in eq 5 for β -GlcNAc. Except for the reducing end residue, for which $\langle ^3J \rangle$ is underestimated by 1.6 Hz, the values computed using eq 5 and the MD conformations are in excellent agreement with values for NMR experiments on hyaluronan oligomers (Table 4). The conformational properties of the acetamido group are seen to be sensitive to the local environment, as was the case for the α - vs β -anomers of GlcNAc. In particular, the reducing-end GlcNAc acetamido group, which has its C1 hydroxyl in the β -anomeric configuration, primarily samples the *cis* conformation, whereas the other three GlcNAc acetamido residues preferentially sample the *trans* conformation, and this conformational difference is reflected in the slightly lower value of $\langle ^3J \rangle$ for GlcNAc1 (Table 4). The key difference between the local environments of the acetamido group for GlcNAc1 vs the three other GlcNAc residues is the presence of the carboxyl group on the preceding GlcUA residue. As GlcNAc1 has no such neighboring residue, its acetamido group cannot act as a hydrogen bond donor to the neighboring carboxyl group, which when formed acts to

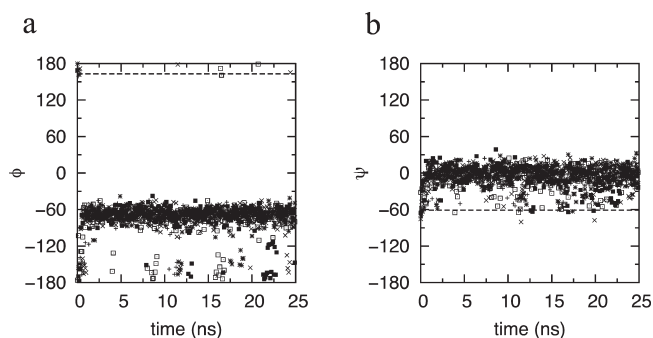


Figure 9. Time-dependent conformational properties of the Neu5Ac α -(2 \rightarrow 3)Gal β glycosidic linkage in sLe^X. Data are shown for the ϕ dihedral angle (a; defined as C1-C2-O_{link}-C3) and the ψ dihedral (b; as C2-O_{link}-C3-H3) for all five 25 ns simulations. Dashed lines indicate values from the reference NMR structure used to seed the simulations.

stabilize the *trans* conformation.⁸² Lacking this stabilization, the GlcNAc1 acetamido group shows conformational behavior similar to that described above for the β -anomer of GlcNAc monosaccharides.

The HA8 ϕ/ψ dihedral distributions from these simulations are consistent with NMR⁸² and X-ray crystallographic⁸³ structures of oligomeric hyaluronan. There is a pronounced global free-energy minimum in each of the distributions for the GlcUA- $\beta(1\rightarrow3)$ -GlcNAc and GlcNAc- $\beta(1\rightarrow4)$ -GlcUA linkages, and for both linkage types, the location of the MD global free-energy minimum coincides with the experimental ϕ/ψ values (Figure 8b,c). While there are two additional local free-energy minima in the case of GlcUA- $\beta(1\rightarrow3)$ -GlcNAc and one additional minimum in the case of GlcUA- $\beta(1\rightarrow4)$ -GlcNAc, these are 2–3 kcal/mol higher in free energy relative to the global minimum and as such correspond to <5% of the sampled conformations. Therefore, the glycosidic linkage parameters that were previously developed using model compound hexopyranose disaccharide analogs lacking hydroxyl or hydroxymethyl groups and validated in the context of hexopyranoses²⁷ do demonstrate transferability to a polymer composed of hexopyranose derivatives.

Conformational Properties of Sialyl Lewis X. Sialyl Lewis X (sLe^X) is a tetrasaccharide carbohydrate moiety of particular importance in molecular recognition and has roles in normal cell function such as leukocyte homing^{84,85} as well as in disease states such as cancer⁸⁶ and chronic inflammatory conditions.⁸⁷ The sLe^X tetrasaccharide consists of Neu5Ac $\alpha(2\rightarrow3)$ Gal $\beta(1\rightarrow4)$ -[Fuc $\alpha(1\rightarrow3)$]GlcNAc β O-R, where “R” indicates linkage of the reducing-end GlcNAc to another moiety. Given the three glycosidic linkages connecting the four component monosaccharides, a significant degree of conformational heterogeneity is, in principle, possible, and pioneering work combined carbohydrate synthesis, NMR spectroscopy, and computational studies to elucidate the structural and conformational properties of sLe^X.^{88–92} Here, using glycosidic dihedral angles from one of these studies,⁸⁹ sLe^X was built, solvated in a truncated octahedron of water molecules extending at least 10 Å in all directions from the solute molecule and with a single neutralizing sodium counterion, briefly minimized and heated with harmonic restraining potentials on sLe^X heavy atom positions, and then, after removal of the positional restraints, simulated for 25 ns with a cutoff value $c = 10$ Å and a time step $dt = 0.002$ ps. Five such

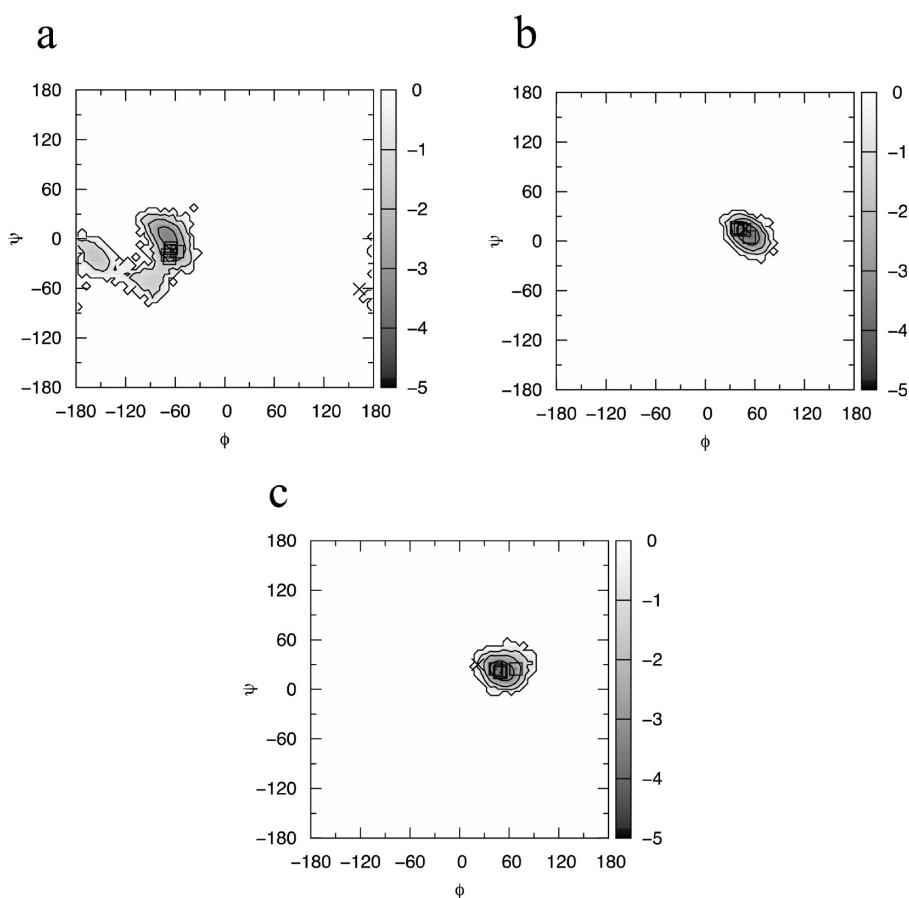


Figure 10. Conformational properties of the sLe^x glycosidic linkages Neu5Ac α (2 \rightarrow 3)Gal β (a; $\phi/\psi = C1-C2-O_{link}-C3/C2-O_{link}-C3-H3$), Gal β (1 \rightarrow 4)GlcNAc (b; $\phi/\psi = H1-C1-O_{link}-C4/C1-O_{link}-C4-H4$), and Fuc α (1 \rightarrow 3)GlcNAc (c; $\phi/\psi = H1-C1-O_{link}-C3/C1-O_{link}-C3-H3$). MD data have been aggregated from the 5–25 ns intervals of five separate simulations and Boltzmann-inverted, with contours every 1 kcal/mol. The 0–5 ns interval from all simulations was excluded to minimize sampling artifacts arising from time-dependent relaxation of the starting conformation. \times 's indicate values from the reference NMR structure used to seed the simulations, and squares are values from sLe^x :protein noncovalent complexes from the PDB (see text).

simulations were done, with random assignment of initial velocities to generate different MD trajectories.

While two of the three sets of glycosidic linkage dihedral angles retained their initial values for all 25 ns in all five trajectories, the third set, namely Neu5Ac α (2 \rightarrow 3)Gal β , was not stable in any of the trajectories. Rather, the ϕ dihedral angle, defined as C1–C2–O_{link}–C3, rapidly relaxed in all cases from the initial value of +163° to one of –70° (Figure 9a). Likewise the ψ dihedral, defined as C2–O_{link}–C3–H3, rapidly relaxed in all cases from the initial value of –61° to one of 0° (Figure 9b). Interestingly, this relaxation was to the region of glycosidic ϕ/ψ space corresponding to that sampled in a previous combined NMR/MD study⁹² that represents one of the pioneering studies on the conformational properties of sLe^x . Importantly, in that study, a 5 ns MD simulation was done *in vacuo* with a dielectric constant of 80 to account for solvent screening of electrostatic effects; this was in contrast to the reference study,⁸⁹ in which *in vacuo* molecular mechanics minimization was used for model refinement.

In the time since the initial work on isolated sLe^x , crystal structures of protein: sLe^x complexes have become available. Searches of the PDB using first the search term “slex” and second the search term “lewis x” yielded six such complexes as of January 2011 (PDB IDs: 1G1R, 1G1T, 2KMB, 2R61, 2RDG,

2Z8L), all of which are noncovalent complexes of sLe^x with sLe^x -binding proteins. For comparison with the reference NMR ϕ/ψ angles, missing hydrogen atoms were assigned to the crystal sLe^x structures using force field geometries, and then each complete crystallographic sLe^x molecule was minimized with harmonic dihedral restraining potentials on heavy-atom ϕ/ψ dihedral angles. Using these optimized crystallographic models, ϕ/ψ angles were noted to be in the same region as the global free-energy minimum in the present MD study for the Neu5Ac α (2 \rightarrow 3)Gal β linkage, as well as for the other two glycosidic linkages in sLe^x (Figure 10). Thus, in accord with early NMR/MD work⁹² and later crystallographic work, each of the three ϕ/ψ angles in sLe^x has a prominent global free-energy minimum, which is preserved in going from aqueous solution to a protein-bound state. Additionally, prior MD and NMR studies^{92,93} have noted that the Neu5Ac α (2 \rightarrow 3)Gal β linkage in unbound sLe^x is not confined to the global minimum, consistent with the present results (Figure 9), whereas upon protein binding this glycosidic linkage does become conformationally constrained,⁹³ consistent with the later crystal structures of sLe^x :protein complexes.

Glycoprotein Systems. Four crystal structures obtained from the PDB database were used to study O and N linkages in a protein environment. These structures were chosen as they

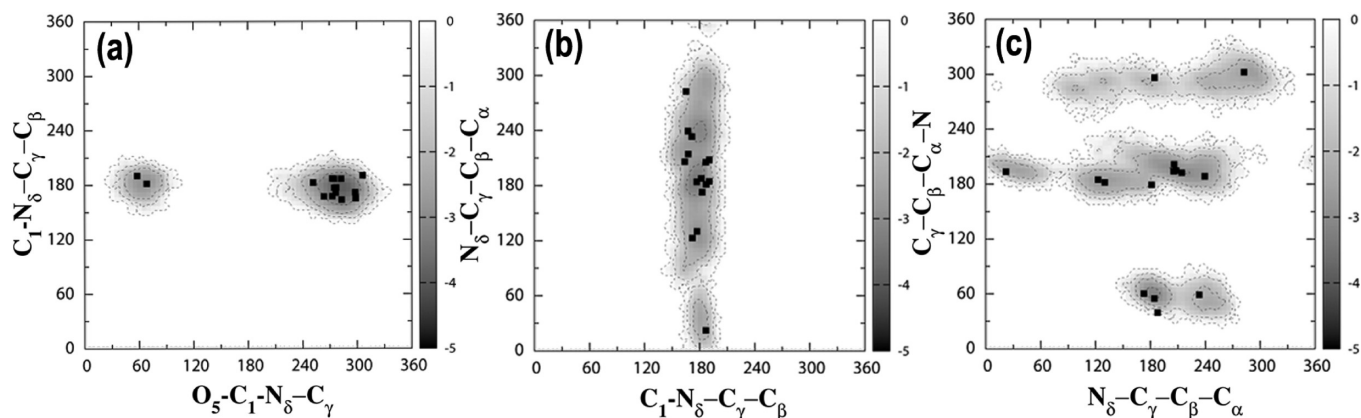


Figure 11. Boltzmann-inverted glycosidic dihedral angle distributions associated with the N-glycan linkage. (a) $O5-C1-N\delta-C\gamma/C1-N\delta-C\gamma-C\beta$ distribution, (b) $C1-N\delta-C\gamma-C\beta/N\delta-C\gamma-C\beta-C\alpha$ distribution, and (c) $N\delta-C\gamma-C\beta-C\alpha/C\gamma-C\beta-C\alpha-N$ distribution. Squares indicate the values observed in the crystallographic structure. Contours are mapped every 1 kcal/mol.

contain multiple glycosylation sites and have been solved at a high resolution (Table S14, Supporting Information). The Reduce software⁹⁴ was used to place missing hydrogen positions and to choose optimal Asn and Gln side chain amide and His side chain ring orientations. Patch residues were used to incorporate disulfide bonds and the O- and N-glycosidic linkages between Ser/Thr or Asn residues and the relevant sugar units. Crystallographic water molecules, counterions, and heteroatoms were included while building the crystal structure for simulations. Scripts obtained from CHARMM-GUI⁹⁵ and modified accordingly were used to set up the simulations, and the CHARMM software was used to solvate each system in a box of dimensions chosen so as to have 10 Å between the protein extremities and the edge of the solvent box (Table S14). Systems were neutralized by adding the appropriate number of counterions and then energy minimized. MD equilibration involved a 100 ps NVT simulation in which harmonic restraints were applied on the protein and the carbohydrate moieties followed by a 200 ps NPT simulation in which all of the restraints were removed. The equilibrated structures were then used for 16 ns production simulations that were performed using NAMD version 2.7b1,⁹⁶ the last 10 ns of which were used for subsequent analysis of ensemble properties.

Analysis of the trajectories revealed a common theme, namely, that the glycan portions of the glycoprotein systems exhibit greater conformational variability than the protein portions. For all systems studied, the overall RMSDs of the complete glycoproteins remain lower than 3 Å for the entire simulation lengths (Figure S8a, Supporting Information). On decomposing the overall RMSD into carbohydrate (Figure S8b) and protein components (Figure S8c), the carbohydrate regions demonstrate high flexibility with deviations as large as 8 Å in some cases, while the underlying protein regions remain very stable, with the RMSD always lower than 2 Å. The high RMSD for the carbohydrate regions is consistent with the high flexibility of carbohydrates, as observed in both NMR and crystallographic studies. In fact, the high conformational variability of carbohydrate regions, combined with the variable glycosylation of identical sites in a sample of a given protein (“microheterogeneity”), is known to hinder crystallographic studies of glycoproteins, posing a barrier to progress for the accumulation of structural data on glycoproteins.^{97,98}

Pooled data from the last 10 ns of the simulation trajectories were used to assess the flexibility of Asn side chains conjugated to glycans as well as the flexibility of the associated glycosidic linkages.

The key observations from these pooled data are that (1) the sampled conformations strongly overlap with crystallographically observed conformations, suggesting correct placement of the molecular mechanics free-energy minima, and (2) there is greater flexibility associated with dihedral atoms exclusively in the Asn side chain as compared to dihedral atoms involved in the glycosidic linkage. With regard to the glycosidic linkage $O5-C1-N\delta-C\gamma/C1-N\delta-C\gamma-C\beta$ dihedrals, two well-defined minima are sampled in the simulations, and both minima are populated in the crystals (Figure 11a). Moving from the glycan toward the protein backbone, the $C1-N\delta-C\gamma-C\beta/N\delta-C\gamma-C\beta-C\alpha$ dihedrals sample a narrow distribution in the $C1-N\delta-C\gamma-C\beta$ coordinate and a broad distribution in the $N\delta-C\gamma-C\beta-C\alpha$ Asn side chain atoms, similar to that seen in crystal structures (Figure 11b). Finally, looking at sampling of the $N\delta-C\gamma-C\beta-C\alpha/C\gamma-C\beta-C\alpha-N$, wherein all eight atoms belong to the Asn side chain, a great deal of flexibility is seen, both in the simulations and in the crystal structures (Figure 11c). Here, the $N\delta-C\gamma-C\beta-C\alpha$ dihedral confers high flexibility to the N linkage. In addition to being consistent with the crystal structures considered here, the varying degrees of conformational flexibility in the simulation data are consistent with a survey of over 500 N-linked glycans in the PDB.⁹⁹ Of particular note is the $C\gamma-C\beta-C\alpha-N$ dihedral, which adopts three well-defined conformations with values of 60°, 180°, and 300° corresponding to the *g+*, *anti*, and *g-* conformational states. The simulation probabilities are *g-* (23%), *anti* (64%), and *g+* (13%), which compares favorably to probabilities from the latter survey of *g-* (18%), *anti* (50%), and *g+* (32%).

Pooled data for O linkages—namely $O5-C1-O1-C\beta/C1-O1-C\beta-C\alpha$ and $C1-O1-C\beta-C\alpha/O1-C\beta-C\alpha-N$ dihedral distributions—are presented in Figure 12. The type of O linkage affected the flexibility of the $C1-O1-C\beta-C\alpha$ dihedral, which was found to be flexible in Ser linkages but more restricted in Thr linkages, with the dihedral sampling conformations around +120° for α -Thr linkages and around +150° for β -Thr linkages. This latter pattern of sampling has also been observed in a combined NMR/MD study of model (α/β)Thr-O-GalNAc diamides.¹⁰⁰ The $O1-C\beta-C\alpha-N$ dihedral is found to adopt three well-defined conformations with values of -60°, ±180°, and +60°, which correspond to the *g-*, *anti*, and *g+* conformational states, which in turn influences the folding-back of the carbohydrate moiety onto the peptide backbone. Additional analysis of the individual linkages revealed $O5-C1-O1-C\beta$ dihedral angle

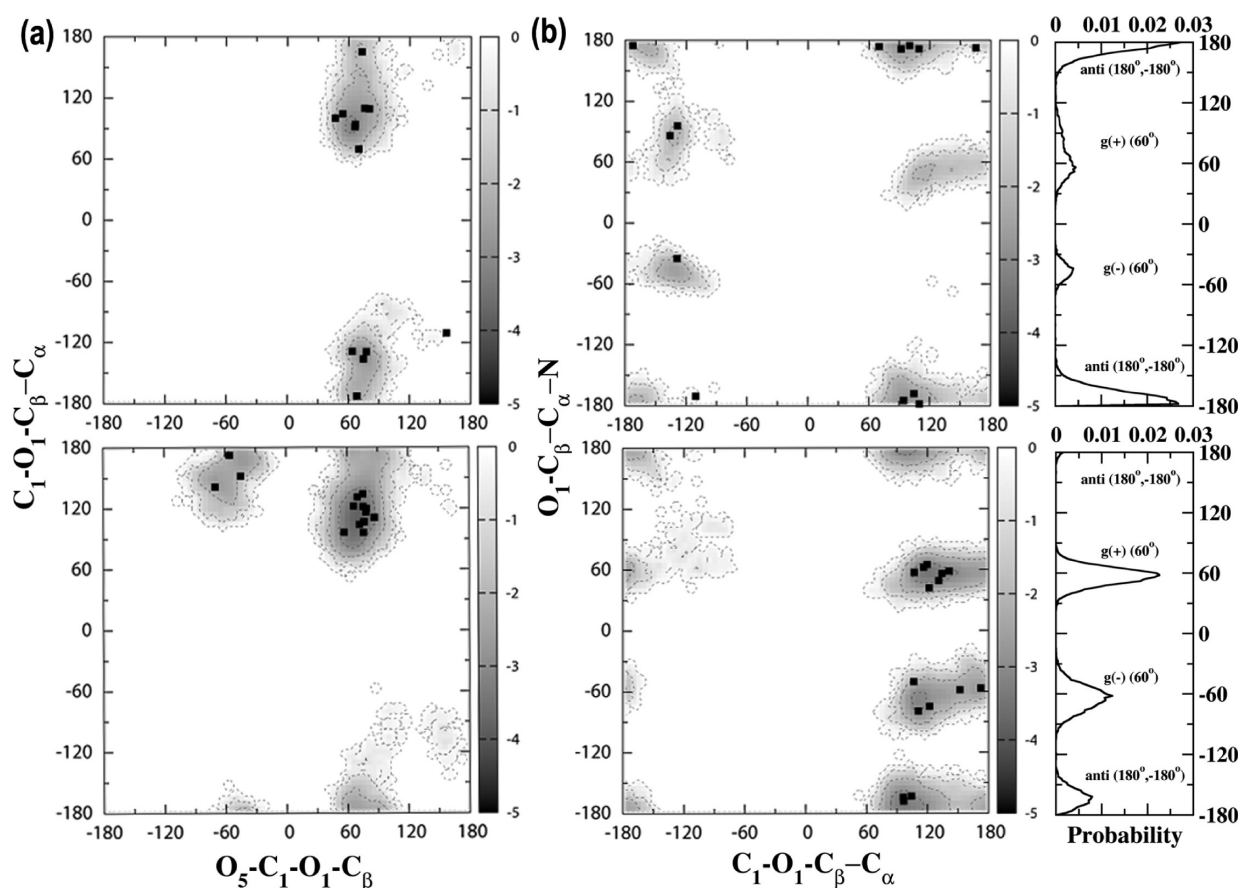


Figure 12. Boltzmann-inverted glycosidic dihedral angle distributions associated with O-glycan linkages. (a) $O_5-C_1-O_1-C_\beta/C_1-O_1-C_\beta-C_\alpha$ distributions and (b) $C_1-O_1-C_\beta-C_\alpha/O_1-C_\beta-C_\alpha-N$ distributions are shown, with data collected from the Ser O-linkages in the top panel and from Thr O-linkages in the lower panel. Squares indicate the values observed for the Ser and Thr O-linkages in the crystallographic structures respectively. The side panel of b contains probability distributions associated with the $O_1-C_\beta-C_\alpha-N$ dihedral angle. Contours are mapped every 1 kcal/mol.

sampling in the region of $+60^\circ$ for the α anomers and -60° for the β anomers, consistent with the exoanomeric effect seen in sugars.¹⁰¹

Lectin–Sucrose Noncovalent Interactions. The designed chimeric cyanovirin-N homologue protein¹⁰² is composed of two domains (A and B), each of which binds one sucrose molecule in sites well separated from each other. Since both X-ray and NMR structures of the complex have been solved,¹⁰² this protein was chosen as a test case for noncovalent protein–carbohydrate interactions. Chain A was chosen out of the two very similar molecules resolved in the crystal asymmetric unit cell of PDB entry 3HP8.¹⁰² The system preparation consisted of adding the three missing N-terminal residues to the protein using the MODELER package,^{103–106} followed by applying the Reduce software⁹⁴ to choose optimal Asn and Gln side chain amide and His side chain ring orientations and the CHARMM software¹⁰⁷ to add missing hydrogens and solvate the system in a rectangular box with dimensions $74 \text{ \AA} \times 53.8 \text{ \AA} \times 52 \text{ \AA}$, chosen to have 10 \AA between the protein extremities and the edge of the solvent box; the net system charge was made neutral by replacing four randomly chosen water molecules with sodium ions. The system was minimized, heated by periodic reassignment of velocities, and equilibrated for 50 ps, all with harmonic restraints on protein and sucrose atoms, after which the system was simulated for 21 ns without restraints, the last 20 ns of which was used to collect data for analysis.

During the 20 ns simulation, the sucrose molecule associated with the A domain (SucA) and the sucrose molecule associated with the B domain (SucB) remained bound to the shallow, surface-exposed sites in domains A and B, respectively. Over the course of the simulation, both molecules sampled only a narrow range of glycosidic ϕ -dihedral values (Figure 13a,b). However, the much broader range of glycosidic ψ -dihedral values spanned by SucA points to greater flexibility of the sucrose molecule bound to domain A. In addition to the conformational region around $\psi = -60^\circ$, which is populated by SucB, SucA also visits regions near $\psi = +60^\circ$. These observations, particularly the alternate $\psi = +60^\circ$ conformational basin populated by SucA, are consistent with the conformational behavior of sucrose in solution, as studied previously.¹⁰⁸ This greater flexibility is mirrored in the higher RMSD of SucA compared to SucB (Figure 13c), despite both molecules remaining bound to their respective pockets. The higher flexibility of SucA observed in the simulation is consistent with experimental data for the system: more NMR resonances are affected in domain A as a result of sucrose binding than in domain B; the average of the SucA atoms' crystallographic B factors is 33.7 \AA^2 vs 18.4 \AA^2 for SucB, and though both are only weakly bound, the experimentally measured apparent binding affinity of SucA is lower than that of SucB ($K_d = 15.2$ and 7.3 mM, respectively).¹⁰²

To better understand the differences between SucA and SucB, the probability of protein–sucrose hydrogen bonds (H bonds)

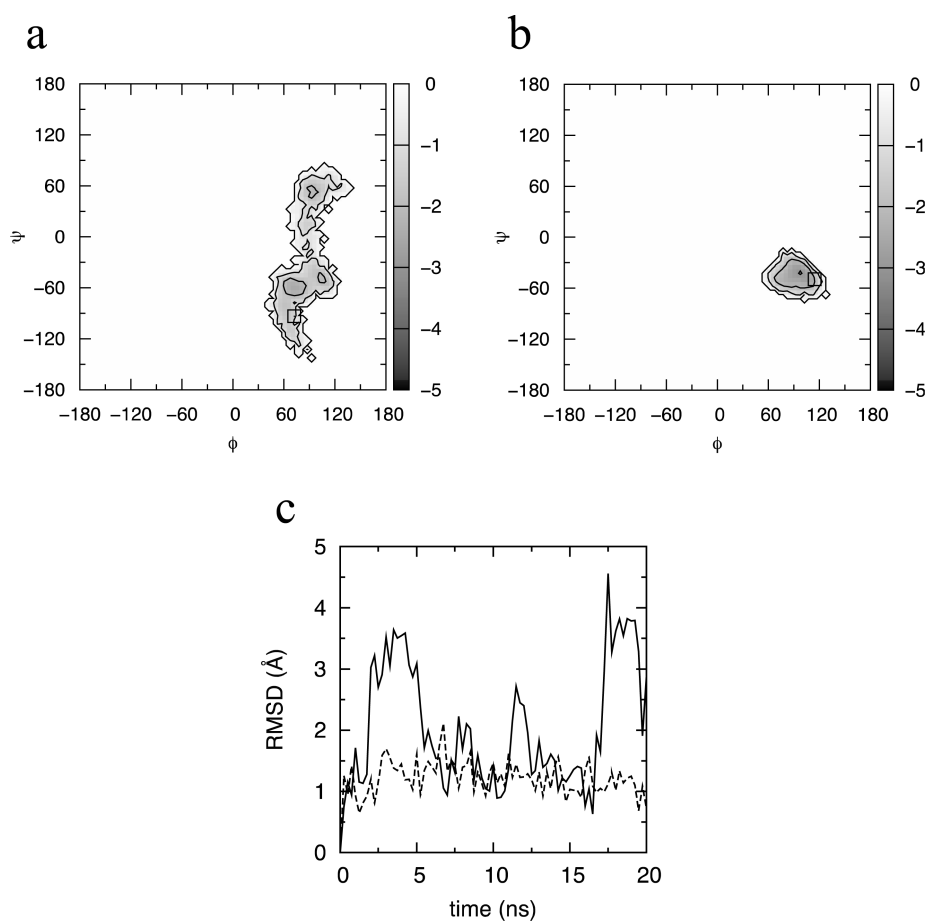


Figure 13. Glycosidic ϕ/ψ ($^g\text{O}_{\text{ring}}-\text{}^g\text{C1}-\text{O}_{\text{link}}-\text{}^f\text{C2}/\text{}^f\text{C1}-\text{O}_{\text{link}}-\text{}^f\text{C2}-\text{}^f\text{O}_{\text{ring}}$, where the superscripts “g” and “f” indicate the glucose and the fructose groups, respectively) dihedral angle distributions and RMSD values for sucrose bound noncovalently to the designed chimeric cyanovirin-N homologue protein. Boltzmann-inverted ϕ/ψ distributions are shown for the sucrose molecule bound to the A domain (a) and the sucrose molecule bound to the B domain (b; contours every 1 kcal/mol), as well as heavy-atom RMSD values for the A-domain and B-domain sucrose molecules with respect to the crystallographic coordinates (c; solid line and dashed line, respectively). Squares indicate the values observed in the crystallographic structure.

was analyzed. The presence of a hydrogen bond was based on a distance cutoff of 2.4 \AA for the acceptor to hydrogen distance. The SucA glucose moiety preserves the H bonds observed in the crystal structure between the C3 hydroxyl group and the Asn99 backbone amide and Gln98 carbonyl oxygen with >95% probability (residue naming is per ref 102). H bonds between the C4 hydroxyl group and Ser2 backbone carbonyl and Ser6 side chain hydroxyl are preserved with >99% probability. A third H bond between the C2 hydroxyl and carbonyl oxygen of Asn99 is preserved with >99% probability. In contrast, the H bonds formed by the fructose moiety are not as highly preserved as the ones formed by the glucose moiety, and this, combined with flexibility about the ψ dihedral, accounts for the flexibility of bound SucA. In particular, the H bond between the C3-hydroxyl and the Asn99 backbone carbonyl oxygen and Asn101 amide is preserved with 56% and 51% probability, respectively, whereas the one between the C4-hydroxyl group and Arg24 backbone carbonyl oxygen is preserved with only 37% occupancy.

As with SucA, H bonds for the glucose moiety of SucB tend to be stable during the course of the simulation. In particular, H bonds between the glucose C4-hydroxyl group and the Asn43 backbone amide and Gln53 carbonyl oxygen are preserved with >90% probability, and the C3-hydroxyl oxygen of the glucose moiety preserves a water mediated H bond with the

Gln53 backbone amide with 38% probability. And while the water mediated hydrogen bond between the C3-hydroxyl hydrogen and Asn54 side-chain carbonyl observed in the crystal is preserved with only 10% occupancy, the glucose ring nonetheless remains firmly bound in its crystallographic conformation. Additionally, unlike SucA, H bonds observed in the crystal between the protein and the SucB fructose moiety are maintained. These include preservation of the hydrogen bonding involving the C3-hydroxyl group and Asn43 carbonyl oxygen and D45 backbone amide with >80% probability. The H bond between the C4 hydroxyl group and carbonyl oxygen of Arg81 is preserved with 91% probability. The preservation of the H-bond network between the fructose moiety of SucB and the protein is consistent with the lesser flexibility about the ψ dihedral for SucB (Figure 13b).

On the basis of these simulations, while binding of the glucose moiety is preserved for sucrose in both binding pockets, subtle structural differences in the binding pockets in the two domains yield a sucrose molecule bound to domain A with a higher degree of flexibility, consistent with NMR, crystallographic, and binding affinity data.¹⁰² These results suggest that the interactions between the carbohydrate and protein aspects of the force field, as well as competition with the solvent, are properly balanced, an outcome of the consistent approach used for the optimization of

the nonbond parameters in the comprehensive CHARMM additive force field for biomolecules.

CONCLUSIONS

The present set of parameters is an important addition to the existing CHARMM carbohydrate force field, as it enables the modeling of common eukaryotic glycans, including glycoproteins. The parametrization has in fact already shown its utility in studying such systems. In one case, simulations were undertaken on a series of compounds containing five different sugars and the dipeptides of Ser and Thr, yielding 14 molecules when the different anomers are taken into account.¹⁰⁹ For eight of these molecules, NMR experimental *J*-coupling and NOE solution data are available,^{100,110,111} and there was overall excellent agreement between the experimental NMR observables and those calculated from simulations using the present force field. In another case, simulations were undertaken on the glycosaminoglycan polymer hyaluronan in a noncovalent complex with the hyaluronan-binding domain of the type I transmembrane protein CD44, resulting in the description of two key monosaccharides in the polymer important for binding as well as a key residue in the protein involved in conformational switching of the hyaluronan-binding site.¹¹² Additional future directions of interest include evaluation of the ring conformational equilibria including the complicated behavior of iduronate,^{113,114} glycosidic conformational transitions that can occur on time scales longer than tens of nanoseconds,¹¹⁵ and the force field description of sulfated and phosphorylated carbohydrates. With regard to this latter direction, work is underway both with regard to parametrization and application.

One consistent trend in the present work is the overestimation of crystal volumes for neutral compounds; this trend is not unexpected given similar results in CHARMM force field models for hexopyranose and furanose monosaccharides,^{24,25} linear sugars and sugar alcohols,²⁶ and disaccharides.^{27,28} One possible explanation is that the highly directional hydrogen bonding in the crystal environment is at odds with the parametrization protocol for hydroxyl groups, which targeted the molecular volumes and heats of vaporization of neat alcohols and therefore is, in a sense, a mean-field approach to developing transferable additive force field parameters. Current work on introducing electronic polarizability into the molecular mechanics framework may help to alleviate this limitation, ideally by yielding a force field where a single set of parameters can yield quantitative results in the gas phase, the crystalline environment, and both aqueous and organic solutions.¹¹⁶

Finally, it is worth noting that much of the present work is transferable to glycans linked to lipids,¹¹⁷ which represent another major class of biomolecules—in addition to proteins—having covalent linkages to carbohydrates. The completion of work presently underway toward this aim will result in an optimized CHARMM additive force field capable of describing the vast majority of heterogeneous biomolecular systems known in eukaryotic biology.

ASSOCIATED CONTENT

S Supporting Information. Model compound and crystalline intramolecular geometries, descriptions of O- and N-linked systems, model compound vibrational frequencies, model compound:water pair interaction geometries, dihedral potential energy scans for

O-linked model compounds, and RMSD analysis of N-linked glycoprotein MD simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: 410/706-7442. Fax: 410/706-5017. E-mail: alex@outerbanks.umaryland.edu.

ACKNOWLEDGMENT

Financial support from the NIH (GM070855, ADM) and University of New England College of Pharmacy startup funds (O.G.) is acknowledged, as is computational support from the Department of Defense and NPACI Alliance, and we thank Drs. Rich Pastor, Richard Venable, and John Brady for helpful discussions.

REFERENCES

- (1) Helenius, A.; Aebi, M. *Annu. Rev. Biochem.* **2004**, *73*, 1019.
- (2) Hakomori, S. *Biochim. Biophys. Acta* **1999**, *1473*, 247.
- (3) Lue, J.; Hsu, M.; Yang, D.; Marx, P.; Chen, Z.; Cheng-Mayer, C. *J. Virol.* **2002**, *76*, 10299.
- (4) Viswanathan, K.; Chandrasekaran, A.; Srinivasan, A.; Raman, R.; Sasisekharan, V.; Sasisekharan, R. *Glycoconj. J.* **2010**, *27*, 561.
- (5) Glennon, T. M.; Zheng, Y. J.; Legrand, S. M.; Shutzberg, B. A.; Merz, K. M. *J. Comput. Chem.* **1994**, *15*, 1019.
- (6) Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraserreid, B. *J. Phys. Chem.* **1995**, *99*, 3832.
- (7) Ott, K. H.; Meyer, B. *J. Comput. Chem.* **1996**, *17*, 1068.
- (8) Senderowitz, H.; Parish, C.; Still, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 2078.
- (9) Reiling, S.; Schlenkrich, M.; Brickmann, J. *J. Comput. Chem.* **1996**, *17*, 450.
- (10) Senderowitz, H.; Still, W. C. *J. Org. Chem.* **1997**, *62*, 1427.
- (11) Durier, V.; Tristram, F.; Vergoten, G. *THEOCHEM* **1997**, *395*, 81.
- (12) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1997**, *18*, 1955.
- (13) Momany, F. A.; Willett, J. L. *Carbohydr. Res.* **2000**, *326*, 194.
- (14) Momany, F. A.; Willett, J. L. *Carbohydr. Res.* **2000**, *326*, 210.
- (15) Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. *J. Comput. Chem.* **2001**, *22*, 1125.
- (16) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10541.
- (17) Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236.
- (18) Kony, D.; Damm, W.; Stoll, S.; van Gunsteren, W. F. *J. Comput. Chem.* **2002**, *23*, 1416.
- (19) Lii, J. H.; Chen, K. H.; Allinger, N. L. *J. Comput. Chem.* **2003**, *24*, 1504.
- (20) Lins, R. D.; Hunenberger, P. H. *J. Comput. Chem.* **2005**, *26*, 1400.
- (21) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622.
- (22) Hansen, H. S.; Hunenberger, P. H. *J. Comput. Chem.* **2011**, *32*, 998.
- (23) Tessier, M. B.; DeMarco, M. L.; Yongye, A. B.; Woods, R. J. *Mol. Simul.* **2008**, *34*, 349.
- (24) Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2008**, *29*, 2543.
- (25) Hatcher, E.; Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2009**, *113*, 12466.

- (26) Hatcher, E. R.; Guvench, O.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2009**, *5*, 1315.
- (27) Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2009**, *5*, 2353.
- (28) Raman, E. P.; Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2010**, *114*, 12981.
- (29) Hatcher, E.; S aw en, E.; Widmalm, G.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2011**, *115*, 597.
- (30) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evansack, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wi orkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
- (31) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400.
- (32) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86.
- (33) MacKerell, A. D., Jr.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105.
- (34) Schlenkrich, M.; Brinkman, J.; MacKerell, A. D., Jr.; Karplus, M. An empirical potential energy function for phospholipids: criteria for parameter optimization and applications. In *Membrane Structure and Dynamics*; Merz, K. M., Roux, B., Eds.; Birkhauser: Boston, 1996; pp 31.
- (35) Feller, S. E.; Gawrisch, K.; MacKerell, A. D., Jr. *J. Am. Chem. Soc.* **2002**, *124*, 318.
- (36) Yin, D. X.; MacKerell, A. D., Jr. *J. Comput. Chem.* **1998**, *19*, 334.
- (37) Feller, S. E.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2000**, *104*, 7510.
- (38) Klauda, J. B.; Brooks, B. R.; MacKerell, A. D., Jr.; Venable, R. M.; Pastor, R. W. *J. Phys. Chem. B* **2005**, *109*, 5300.
- (39) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2010**, *31*, 671.
- (40) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (41) MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The energy function and its parameterization with an overview of the program. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 1; pp 271.
- (42) Guvench, O.; MacKerell, A. D., Jr. Comparison of protein force fields for molecular dynamics simulations. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Humana Press, Inc.: Totowa, NJ, 2008; pp 63.
- (43) MacKerell, A. D., Jr.; Wi orkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946.
- (44) Feller, S. E.; Yin, D. X.; Pastor, R. W.; MacKerell, A. D., Jr. *Biophys. J.* **1997**, *73*, 2269.
- (45) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D., Jr.; Pastor, R. W. *J. Phys. Chem. B* **2010**, *114*, 7830.
- (46) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U. K., 1987.
- (47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (48) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. *J. Phys. Chem.* **1994**, *98*, 2198.
- (49) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (50) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667.
- (51) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
- (52) Hockney, R. W. The potential calculation and some applications. In *Methods in Computational Physics*; Alder, B., Fernbach, S., Rotenberg, M., Eds.; Academic Press: New York, 1970; Vol. 9; pp 136.
- (53) Nos e, S. *Mol. Phys.* **1984**, *52*, 255.
- (54) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.
- (55) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613.
- (56) Allen, F. H. *Acta Crystallogr. Sect. B* **2002**, *58*, 380.
- (57) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T., Jr.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, K.; Kitao, O.; Nakai, H.; Klene, M.; Li, T. W.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Revision B.04 ed.; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (58) M oller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (59) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (60) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.
- (61) Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. A* **2006**, *110*, 9934.
- (62) Woodcock, H. L.; Moran, D.; Pastor, R. W.; MacKerell, A. D., Jr.; Brooks, B. R. *Biophys. J.* **2007**, *93*, 1.
- (63) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502.
- (64) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. *J. Am. Chem. Soc.* **1979**, *101*, 2550.
- (65) Guvench, O.; MacKerell, A. D., Jr. *J. Mol. Model.* **2008**, *14*, 667.
- (66) MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584.
- (67) MacKerell, A. D., Jr.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 10559.
- (68) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2005**, *109*, 18988.
- (69) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2007**, *3*, 1120.
- (70) *Essentials of Glycobiology*, 2nd ed.; Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Marilynn, E. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2009.
- (71) *Foye's Principles of Medicinal Chemistry*, 6th ed.; Lemke, T. L., Williams, D. A., Eds.; Lippincott, Williams, and Wilkins: Baltimore, MD, 2008.
- (72) Varki, A. *Nature* **2007**, *446*, 1023.
- (73) Spiro, R. G. *Glycobiology* **2002**, *12*, 43R.
- (74) Strous, G. J.; Dekker, J. *Crit. Rev. Biochem. Mol. Biol.* **1992**, *27*, 57.
- (75) Zachara, N. E.; Hart, G. W. *Chem. Rev.* **2002**, *102*, 431.
- (76) Dwek, R. A. *Chem. Rev.* **1996**, *96*, 683.
- (77) Hart, G. W.; Brew, K.; Grant, G. A.; Bradshaw, R. A.; Lennarz, W. J. *J. Biol. Chem.* **1979**, *254*, 9747.
- (78) Bause, E. *Biochem. J.* **1983**, *209*, 331.
- (79) Mobli, M.; Almond, A. *Org. Biomol. Chem.* **2007**, *5*, 2243.
- (80) Toole, B. P. *Nat. Rev. Cancer* **2004**, *4*, 528.
- (81) Almond, A. *Cell. Mol. Life Sci.* **2007**, *64*, 1591.
- (82) Almond, A.; DeAngelis, P. L.; Blundell, C. D. *J. Mol. Biol.* **2006**, *358*, 1256.
- (83) Banerji, S.; Wright, A. J.; Noble, M.; Mahoney, D. J.; Campbell, I. D.; Day, A. J.; Jackson, D. G. *Nat. Struct. Mol. Biol.* **2007**, *14*, 234.
- (84) Kannagi, R. *Curr. Opin. Struct. Biol.* **2002**, *12*, 599.
- (85) Sperandio, M. *FEBS J.* **2006**, *273*, 4377.
- (86) Cazet, A.; Julien, S.; Bobowski, M.; Krzewinski-Recchi, M. A.; Harduin-Lepers, A.; Groux-Degroote, S.; Delannoy, P. *Carbohydr. Res.* **2010**, *345*, 1377.
- (87) Romano, S. J. *Treat. Respir. Med.* **2005**, *4*, 85.
- (88) Berg, E. L.; Robinson, M. K.; Mansson, O.; Butcher, E. C.; Magnani, J. L. *J. Biol. Chem.* **1991**, *266*, 14869.

- (89) Lin, Y. C.; Hummel, C. W.; Huang, D. H.; Ichikawa, Y.; Nicolaou, K. C.; Wong, C. H. *J. Am. Chem. Soc.* **1992**, *114*, 5452.
- (90) Ball, G. E.; O'Neill, R. A.; Schultz, J. E.; Lowe, J. B.; Weston, B. W.; Nagy, J. O.; Brown, E. G.; Hobbs, C. J.; Bednarski, M. D. *J. Am. Chem. Soc.* **1992**, *114*, 5449.
- (91) Ichikawa, Y.; Lin, Y. C.; Dumas, D. P.; Shen, G. J.; Garcia-Junceda, E.; Williams, M. A.; Bayer, R.; Ketcham, C.; Walker, L. E. *J. Am. Chem. Soc.* **1992**, *114*, 9283.
- (92) Rutherford, T. J.; Spackman, D. G.; Simpson, P. J.; Homans, S. W. *Glycobiology* **1994**, *4*, 59.
- (93) Cooke, R. M.; Hale, R. S.; Lister, S. G.; Shah, G.; Weir, M. P. *Biochemistry* **1994**, *33*, 10591.
- (94) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999**, *285*, 1735.
- (95) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008**, *29*, 1859.
- (96) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781.
- (97) Lutteke, T.; Frank, M.; von der Lieth, C. W. *Carbohydr. Res.* **2004**, *339*, 1015.
- (98) Chang, V. T.; Crispin, M.; Aricescu, A. R.; Harvey, D. J.; Nettleship, J. E.; Fennelly, J. A.; Yu, C.; Boles, K. S.; Evans, E. J.; Stuart, D. I.; Dwek, R. A.; Jones, E. Y.; Owens, R. J.; Davis, S. J. *Structure* **2007**, *15*, 267.
- (99) Petrescu, A. J.; Milac, A. L.; Petrescu, S. M.; Dwek, R. A.; Wormald, M. R. *Glycobiology* **2004**, *14*, 103.
- (100) Corzana, F.; Busto, J. H.; Jimenez-Oses, G.; Garcia de Luis, M.; Asensio, J. L.; Jimenez-Barbero, J.; Peregrina, J. M.; Avenoza, A. *J. Am. Chem. Soc.* **2007**, *129*, 9458.
- (101) Rao, V. S. R.; Qasba, P. K.; Balaji, P. V.; Chandrasekaran, R. *Conformation of Carbohydrates*; Harwood Academic Publishers: Amsterdam, 1998.
- (102) Koharudin, L. M.; Furey, W.; Gronenborn, A. M. *Proteins* **2009**, *77*, 904.
- (103) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779.
- (104) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291.
- (105) Fiser, A.; Do, R. K. G.; Sali, A. *Protein Sci.* **2000**, *9*, 1753.
- (106) Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A. *Current Protocols in Bioinformatics*; Wiley: New York, 2006; Chapter 5, Unit 56.
- (107) Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
- (108) Raman, E. P.; Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2010**, *114*, 12981.
- (109) Mallajosyula, S. S.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2011**, DOI:10.1021/jp203695t.
- (110) Corzana, F.; Busto, J. H.; Engelsen, S. B.; Jimenez-Barbero, J.; Asensio, J. L.; Peregrina, J. M.; Avenoza, A. *Chemistry* **2006**, *12*, 7864.
- (111) Fernandez-Tejada, A.; Corzana, F.; Busto, J. H.; Jimenez-Oses, G.; Jimenez-Barbero, J.; Avenoza, A.; Peregrina, J. M. *Chemistry* **2009**, *15*, 7297.
- (112) Jamison, F. W., 2nd; Foster, T. J.; Barker, J. A.; Hills, R. D., Jr.; Guvench, O. *J. Mol. Biol.* **2011**, *406*, 631.
- (113) Babin, V.; Sagui, C. *J. Chem. Phys.* **2010**, *132*, 104108.
- (114) Sattelle, B. M.; Hansen, S. U.; Gardiner, J.; Almond, A. *J. Am. Chem. Soc.* **2010**, *132*, 13132.
- (115) Peric-Hassler, L.; Hansen, H. S.; Baron, R.; Hunenberger, P. H. *Carbohydr. Res.* **2010**, *345*, 1781.
- (116) Lopes, P. E.; Roux, B.; Mackerell, A. D., Jr. *Theor. Chem. Acc.* **2009**, *124*, 11.
- (117) Abel, S.; Dupradeau, F. Y.; Raman, E. P.; MacKerell, A. D., Jr.; Marchi, M. *J. Phys. Chem. B* **2011**, *115*, 487.

Evaluation of DNA Force Fields in Implicit Solvation

Thomas Gaillard^{†,‡} and David A. Case^{*,†}

[†]BioMaPS Institute for Quantitative Biology, Rutgers—The State University of New Jersey, Piscataway, New Jersey 08854-8087, United States

[‡]Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France

ABSTRACT: DNA structural deformations and dynamics are crucial to its interactions in the cell. Theoretical simulations are essential tools used to explore the structure, dynamics, and thermodynamics of biomolecules in a systematic way. Molecular mechanics force fields for DNA have benefited from constant improvements during the past decades. Several studies have evaluated and compared available force fields when the solvent is modeled by explicit molecules. On the other hand, few systematic studies have assessed the quality of duplex DNA models when implicit solvation is employed. The interest in an implicit modeling of the solvent consists of the important gain in simulation performance and conformational sampling speed. In this study, respective influences of the force field and the implicit solvation model choice on DNA simulation quality are evaluated. To this end, extensive implicit solvent duplex DNA simulations are performed, attempting to reach both conformational and sequence diversity convergence. Structural parameters are extracted from simulations and statistically compared to available experimental and explicit solvation simulation data. Our results quantitatively expose the respective strengths and weaknesses of the different DNA force fields and implicit solvation models studied. This work can lead to the suggestion of improvements to current DNA theoretical models.

INTRODUCTION

Deoxyribonucleic acid (DNA) is the molecule carrying genetic information, which is translated into proteins by living cells, using a code almost universally conserved among living systems. DNA is involved in many dynamic processes like gene expression and regulation or genetic replication and recombination. DNA molecule structural plasticity is critical in many of these mechanisms, including protein or cofactor recognition, chromatin remodeling, and transcription. Understanding DNA structural deformations and dynamics is thus essential to the study of important components of living systems.

Experimental techniques like X-ray crystallography, NMR spectroscopy, or single molecule experiments greatly contribute to our increasing knowledge of DNA molecule properties. Theoretical simulations are complementary techniques, useful in the investigation of structure, dynamics, and thermodynamics of biomolecules in a systematic way, in particular when experimental data are not available or accessible. The first molecular dynamics simulations of DNA were performed in 1983^{1,2} in vacuo and two years later with explicit water molecules and counterions.³ DNA molecular mechanics models have thus benefited from almost three decades of force field and algorithmic improvements^{4–12} and have been applied to a wide variety of questions.^{13–21}

A number of systematic studies have attempted to evaluate DNA molecular mechanics models.^{22–29} Only the most recent works are described here. In a collaborative work initiated by the “Ascona B-DNA Consortium”, 136 unique tetranucleotide duplexes were studied by molecular dynamics. Simulations were performed over 15 ns, using the AMBER parm94 force field, with an explicit solvation model, on 39 oligomers of 15 bp length containing the tetranucleotides. Methodological questions are addressed in a first paper,²⁵ and sequence context effects on the DNA structure are discussed in a second paper.²⁶ The authors observed structural substates distinct from the B form. These states are largely controlled by the backbone conformation,

which in turn is strongly correlated to helicoidal parameters. The effect of flanking base pairs on a dinucleotide was found to be limited for YpR steps, although they are more flexible, and more significant for RpR and RpY steps, which are more rigid. Another study by Fujii et al.²⁷ concentrated on the sequence dependence of DNA deformability. Molecular dynamics simulations of 10 ns length in an aqueous environment were performed on 12 bp duplexes, comprising the 136 tetramers sandwiched between CGCG sequences. The AMBER parm99 force field was used. The authors showed that the deformability of dimeric steps is consistent with crystal structure data and can be affected by flanking base pairs. Orozco and co-workers²⁸ conducted a systematic study of B-DNA flexibility with CHARMM all27 and AMBER parmbsc0 force fields. They performed 100 ns molecular dynamics simulations in an explicit solvent environment on four different 18 bp sequences, chosen to contain the 10 unique dinucleotide base pair steps. This study highlighted some differences between parmbsc0 and all27 force fields, yet the authors concluded that today DNA force fields have matured, leading to a consensus view of B-DNA flexibility. A continuation of the Ascona consortium effort was recently published.²⁹ This work investigated nearest-neighbor effects on B-DNA base pairs and base pair steps. A number of 39 oligomers containing all unique tetranucleotides, each of 18 base pair length, were subjected to 50 or 100 ns molecular dynamics simulations in explicit solvation, with the parmbsc0 force field. The authors concluded that simulations were converged in terms of most B-DNA conformational properties. Sequence effects were found to be small at the base or base pair level, more important at the tri- or tetranucleotide level, and could still be significant beyond the nearest-neighbor level.

Received: June 8, 2011

Published: August 24, 2011

Solvent modeling is critical in DNA simulations. With an improper solvation model, DNA strand separation can easily be observed. While explicit modeling of solvent molecules provides the most realistic results, the cost of such approaches is high, as an important portion of computer time is spent in simulating bulk solvent molecule dynamics. Implicit solvation models attempt to represent solvent as a continuous medium and capture effects on the solute in an average way. The interest of implicit solvation models lies in reduced computational cost, reduced statistical error, faster conformational sampling, clearer physical insights, and the possibility of modeling big systems like chromatin or systems with large conformational changes. An important class of implicit solvation models includes the methods based on continuum electrostatics where the electrostatic component of the solvation free energy is obtained by solving the Poisson–Boltzmann equation (PB). The generalized Born model³⁰ (GB) is a computationally efficient approximation of PB, suitable for molecular dynamics simulations. Several GB variants have been implemented in molecular mechanics programs, mostly differing by the way effective Born radii are calculated. Some GB methods are based on the pairwise descreening formulation of Hawkins and co-workers,^{31–39} others approximate the solute volume by overlapping Gaussian functions,^{40,41} while others calculate Born radii by analytical volume integration.^{42–44} Performance comparisons of several GB implementations are available.^{45–47}

GB implicit solvation models have been widely applied,^{48–51} in particular to proteins and to a lesser extent to nucleic acids.^{36,52–56} Srinivasan et al.⁵² conducted PB and GB calculations on snapshots extracted from explicit solvent molecular dynamics simulations of DNA, phosphoramidate-modified DNA, and RNA decamers in the A or B form. Free energy estimations confirmed that the B form is preferred for DNA and the A form for RNA and phosphoramidate-modified DNA and that salt inclusion modestly favors the A form. In another work, Srinivasan et al.⁵³ performed PB and GB calculations on multiple conformations of proteins, 10–12 bp DNA duplexes, and a RNA hairpin. Good agreement is found between GB and PB solvation energies and salt contributions. Tsui and Case^{36,54} have simulated 10 bp DNA and RNA duplexes in the A and B forms using molecular dynamics with GB implicit solvation and obtained good agreement with explicit solvent simulations in terms of both structures and energetics. The authors also showed that DNA in the A form converges to the B form about 20 times faster with GB than with explicit solvation. Chocholoušová and Feig⁵⁵ have performed molecular dynamics simulations with GB implicit solvation on the DNA Drew–Dickerson dodecamer and a protein–DNA complex. The authors demonstrated that stable and realistic DNA simulations can be obtained, agreement being closer to either explicit solvation results or experimental data, depending on the atomic radii set chosen. Partial transitions to A-DNA could be observed when 1 M salt was added. Molecular dynamics simulations with GB implicit solvation have also been applied to bigger nucleic acid assemblies such as nucleosomal DNA.⁵⁶ To our knowledge, there is no systematic evaluation of DNA simulations in implicit solvation published to date.

In this work, we intend to evaluate DNA molecular mechanics models with generalized Born implicit solvents against explicit solvation results and experimental data. Different force fields and implicit solvation methods are implemented and customarily used with different molecular mechanics programs, like CHARMM or AMBER. We need to be able to use any combination of force field and implicit solvation variant to investigate the respective

influence of both on results. We concentrate on two recent and widely used DNA force fields (AMBER parmbsc0 and CHARMM all27) and five generalized Born variants (GBHCT, GBOBC1, GBOBC2, and GBn implemented in AMBER and GBMV implemented in CHARMM). Molecular dynamics simulations are conducted on several DNA duplex sequences, and care is taken that both simulation length and sequence diversity convergence are achieved. Several structural parameters are then measured and statistically compared to both explicit solvation results and experimental data.

METHODS

Starting Structures. DNA duplexes of 12 base pair length with different sequences corresponding to the 5′-CGCGWXYZCGCG-3′/5′-CGCGZYXWCGCG-3′ pattern were studied, where WXYZ/ZYXW are tetramer duplexes located at the center and flanked by C–G base pairs. The overline symbol designates the complementary nucleotide. There are 256 combinations of nucleotide tetramers but only 136 unique tetramer duplexes (for example, AAAA/TTTT and TTTT/AAAA cannot be distinguished). Starting coordinates were built with the NAB program distributed with the AMBER suite,⁵⁷ using Arnott B-DNA fiber diffraction data.⁵⁸

AMBER prmtop and inpcrd files were generated with the leap program, using the parmbsc0 force field.¹² CHARMM psf and crd files were generated with the CHARMM program⁵⁹ using the all27¹¹ force field. “CHAMBER” (CHARMM force field with the AMBER program) prmtop and inpcrd files were generated with the CHAMBER conversion program⁶⁰ distributed with the AMBER suite, from CHARMM psf and crd files, and the all27 rtf and prm files. Differences between CHAMBER and CHARMM energies in vacuo with an infinite cutoff were less than 5×10^{-4} kcal mol⁻¹. “AMBARM” (AMBER force field with the CHARMM program) psf and crd files were generated with the CHARMM program compiled with the AMBER keyword to be consistent with AMBER conversion factors and using the parmbsc0 force field converted to CHARMM format by Jeff Klauda (<http://terpconnect.umd.edu/jbklauda/research/download.html>). Differences between AMBARM and AMBER energies in vacuo with an infinite cutoff were less than 3×10^{-3} kcal mol⁻¹.

Molecular Dynamics. Molecular dynamics simulations with AMBER and CHARMM programs were done with a generalized Born implicit solvation model, using an infinite cutoff for non-bonded interactions, dielectric constants of 78.5 for the solvent and 1 for DNA interior, and an inverse Debye–Hückel length of 0.1 Å⁻¹ to account for the 0.1 M salt concentration. Minimization was first performed over 200 steps with 5 kcal mol⁻¹ Å⁻² harmonic restraints on non-hydrogen atoms. Langevin dynamics was then carried out using a heat bath at 300 K, a collision frequency of 5 ps⁻¹, a 1 fs time step, and the SHAKE algorithm was used to restrain the elongation of bonds involving hydrogen atoms. Three equilibration phases of 50 ps were performed with decreasing harmonic restraints on non-hydrogen atoms of 5, 1, and 0.1 kcal mol⁻¹ Å⁻², respectively. The production phase of 5 ns was then run without restraints. Simulations with the AMBER program employed GBHCT,^{31,32} GBOBC1,^{37,38} GBOBC2,³⁸ or GBn³⁹ generalized Born model variants. Simulations with the CHARMM program employed the GBMV implicit solvation model,^{42,43} with the $1/r^7$ Coulomb field correction and parameters $C_0 = -0.1$, $C_1 = 0.9$, $\beta = -12$, and $S_0 = 0.65$, as in Chocholoušová and Feig.^{55,61} Recommended atomic radii sets for each GB model variant were used (mbondi³⁶ for GBHCT,

mbondi³⁸ for GBOBC1 and GBOBC2, bondi⁶² for GBn, and van der Waals radii for GBMV).

Analysis. Structural data were collected from the 5 ns molecular dynamics trajectories of the 136 CGCGWXYZCGCG duplex sequences, each trajectory containing 5,000 frames. Combinations of two force fields (AMBER parmbsc0 and CHARMM all27) and five GB implicit solvation models (AMBER GBHCT, GBOBC1, GBOBC2, GBn, and CHARMM GBMV) led to a total simulation time of 6.8 μ s and a number of 6.8×10^6 frames.

Trajectories were analyzed with the ptraj program of the AMBER suite⁵⁷ for root-mean-square deviations (RMSD) calculations, and structural parameters⁶³ were evaluated with the 3DNA program.^{64,65} Parameter measurements were done on the central XY/YX dimer and included base pair step parameters (shift, slide, rise, tilt, roll, twist), groove widths (minor and major), sugar conformational parameters (amplitude and phase angle of ring pseudorotation), backbone and glycosidic torsion angles ($\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi$). Only structures with correct Watson–Crick pairing at the central dimer, as defined by 3DNA, were considered.

The data set was then symmetrized by duplicating each entry with parameter values that would have been measured if the other strand was chosen as the “leading strand”. This is necessary because some structural parameters (shift and tilt) have their sign dependent on the choice of the duplex leading strand, which is arbitrary. Structural parameter global averages, standard deviations, and probability density functions were then calculated over the whole data set, using a weighting to correct for sequence composition bias and ensure equal representation to each of the 10 unique dimer steps (AA/TT, AT/AT, AG/CT, AC/GT, TA/TA, TG/CA, TC/GA, GG/CC, GC/GC, CG/CG). Dimer level averages, standard deviations, and probability density functions were also calculated over subsets corresponding to each of the 10 unique dimer steps. The probability density function was approximated by a kernel density estimation, using a Gaussian function as the kernel. Periodicity was taken into account when calculating statistics of angular parameters.

Experimental Data. Experimental structural data were extracted from the Protein Data Bank⁶⁶ and Nucleic Acid Database.⁶⁷ The 3DNA Landscapes platform⁶⁸ (<http://3dnascapes.rutgers.edu>) was employed to obtain PDB/NDB codes of DNA-containing structures. A list of X-ray structures with a resolution better than 2 Å was retrieved as well as a separate list of NMR structures. Structures were then downloaded and analyzed with the 3DNA program. Unwanted structures were discarded according to the following criteria: dimer steps retained in the data set were those containing only A, T, G, or C nucleotides; with correct Watson–Crick pairing; flanked by A, T, G, or C nucleotides; and belonging to a right-handed duplex structure. Further filtering excluded A-DNA ($z_p > 1.5$ Å) and TA-DNA ($z_p(h) > 4.0$ Å) structures, as in Lu and Olson.⁶⁴ The number of dimer steps available in the experimental data set is given in Table 1. Statistical analysis of experimental data was conducted as for implicit solvation MD data.

Explicit Solvation Simulation Data. Structural data were also obtained from molecular dynamics simulations in explicit solvation performed by Orozco and co-workers.²⁸ The 100 ns trajectories of four different sequences simulated with both CHARMM all27 and AMBER parmbsc0 force fields were downloaded (<http://mmb.pcb.ub.es/raist/CONSENSUS/>) and uncompressed, as explained in the article. Each of the eight trajectories contained about 10^5 frames, of which one every 10 frames was extracted and analyzed with the 3DNA program. Every dimer step with correct

Table 1. Amount of Dimer Steps Available in Experimental Data^a

exptl.	NMR	X-ray
all	2788	3018
AA	388	476
AT	200	273
AG	338	270
AC	343	303
TA	172	187
TG	351	341
TC	353	420
GG	249	266
GC	206	238
CG	188	244

^a NMR and X-ray (resolution better than 2 Å) structures are extracted from the Protein Data Bank. Dimer steps retained in the data set were those containing only A, T, G, or C nucleotides, with correct Watson–Crick pairing, flanked by A, T, G, or C nucleotides, and corresponding to a right-handed B-DNA duplex structure.

Watson–Crick pairing and not at the extremities of the duplex to avoid end effects was kept in the data set. Statistical analysis of the data set was then performed as for implicit solvation MD data.

RESULTS

DNA structural parameters statistics, obtained from molecular dynamics simulations with different force fields and generalized Born implicit solvation methods, are presented and compared to experimental X-ray and NMR data, as well as to published explicit solvation simulation data. The validity of simulations and methodological robustness are first discussed. A detailed examination of selected structural parameter statistics is then provided.

Validity of Simulations. *Watson–Crick Pairing.* The Watson–Crick pairing percentage was calculated for each simulation as a measure of DNA duplex structure conservation (Table 2). This was obtained as the ratio of frames for which a correct pairing of the central dimer was detected by the 3DNA program to the total number of frames simulated. Percentages are then averaged over the 136 sequences simulated with each force field and GB method combination. Percentages calculated over subsets corresponding to the 10 different central dimer types are also provided. An examination of the trajectories shows that strands get irreversibly separated after about 500 ps of simulation when the GBn method is employed, thus explaining the low percentages (around 10%) found with GBn. This method is indeed not recommended for nucleic acid simulations by its authors³⁹ and will not be considered in analyses. In all other simulations, Watson–Crick pairing percentages were always higher than 98.5% on average. The AMBER parmbsc0 force field gave slightly higher percentages than CHARMM all27 in general. GBMV gave the best overall pairing conservation among GB methods, while GBOBC1 had the lowest percentages. Limited variations were observed between different types of central dimer.

Root Mean Square Deviations. Root mean square deviations of heavy atoms with respect to the initial conformation were calculated for the different simulations, taking into account either the whole duplex or only the central dimer step (Table 3). The RMSD values given are obtained by conformational averaging over the course of simulations and sequential averaging over the

Table 2. Watson–Crick Pairing Percentages of the Central Dimer in MD Simulations^a

FF	parmbc0					all27				
	solv.	GBHCT	GBOBC1	GBOBC2	GBn	GBMV	GBHCT	GBOBC1	GBOBC2	GBn
all	99.73	99.72	99.89	14.99	99.99	99.42	98.66	99.34	7.95	99.70
AA	99.38	99.76	99.95	4.58	100.00	99.92	99.27	99.71	4.60	99.99
AT	99.95	99.76	99.96	4.15	100.00	99.91	98.78	99.57	4.40	99.92
AG	99.83	99.79	99.87	12.21	99.99	99.56	99.03	99.49	8.56	99.61
AC	99.79	99.60	99.89	12.75	100.00	99.22	98.13	98.99	6.89	99.11
TA	99.63	99.70	99.99	3.29	100.00	99.57	99.29	99.54	4.92	99.73
TG	99.69	99.67	99.77	11.70	99.99	99.50	98.70	99.47	8.19	99.68
TC	99.90	99.78	99.85	11.69	99.99	99.61	99.14	99.67	7.18	99.98
GG	99.74	99.73	99.86	33.50	99.95	98.52	97.40	98.61	10.21	99.39
GC	99.88	99.86	99.97	27.41	99.99	99.33	98.86	99.40	14.32	99.89
CG	99.50	99.58	99.83	30.76	99.99	99.18	98.21	99.01	11.42	99.96

^a Percentages are obtained as the ratio of frames with correct Watson–Crick pairing at the central dimer, as measured by the 3DNA program, to the total number of trajectory frames. They are calculated for all frames as well as over subsets corresponding to each of the 10 unique dimer steps.

Table 3. Heavy-Atom Root Mean Squared Deviation (Å) Averages with Respect to Initial Conformation^a

FF	GB	duplex	central dimer
parmbc0	GBHCT	3.45	1.13
parmbc0	GBOBC1	4.26	1.21
parmbc0	GBOBC2	4.08	1.17
parmbc0	GBn	15.89	9.41
parmbc0	GBMV	2.85	0.92
all27	GBHCT	2.68	0.86
all27	GBOBC1	3.82	0.96
all27	GBOBC2	3.58	0.91
all27	GBn	17.77	11.43
all27	GBMV	2.98	0.77

^a RMSDs are calculated on the whole 12 bp duplex as well as on the central dimer only and averaged over trajectories of the 136 sequences simulated.

136 sequences simulated with each force field and GB variant. Except GBn, for which strand separation occurs, RMSD calculated on the whole 12 bp duplex ranged from 2.7 to 4.3 Å on average, and from 0.8 to 1.2 Å when only the central dimer was considered. Slightly higher values are in general obtained with the parmbc0 force field. The GB method choice influences RMSD in a consistent way between both force fields, namely, GBOBC1 > GBOBC2 > GBHCT ≈ GBMV. The highest RMSD was obtained with GBOBC1 and then GBOBC2, and the lowest values were obtained with the GBMV method with both force fields, or with the all27+GBHCT combination. It should be noted that structural measurements in analyses are done exclusively on the central dimer; results will thus not be directly affected by deformations at the extremities of the duplex. Another remark is that initial structures are not necessarily considered as a reference in this work, where we assess the validity of GB simulations with respect to explicit solvation simulation and experimental data. Initial structures are indeed ideal B-DNA structures based on fiber diffraction data, which are not taking into account sequence effects on backbone conformation.

Convergence and Robustness. Convergence and robustness of the results were assessed by several preliminary tests, employing energy minimizations or molecular dynamics. It was found

that DNA structural parameters were mostly converged with respect to simulation time (5 ns), sequence diversity (136 unique duplex tetramers), and duplex length (12 bp). Results were in general not significantly sensitive to the choice of the starting conformation, salt concentration in the reasonable range, or GB radii set.

Structural Parameters. Structural parameter averages are given in Table 4. Base pair step, groove width, backbone and glycosidic torsions, and sugar pucker parameters are discussed in detail, on the basis of a comparison of averages and whole probability distributions. Most of the time, distributions are close to Gaussian shape, and averages are a good approximation of the maximum of probability. As a general pattern, we first compare NMR to X-ray data, then force fields to experimental data, and finally GB implicit solvation methods to explicit solvation.

Base Pair Step Parameters. Base pair step parameters describe the geometry of one base pair with respect to the next one. There are three translational base pair step parameters (shift, slide, rise) and three rotational parameters (tilt, roll, twist). The shift and tilt parameters, presenting a sign inversion depending on the choice of the leading strand of the duplex, have an average value of zero by construction and will not be discussed in detail here.

Slide. Slide distributions are given in Figure 1. The slide average appears noticeably lower in NMR data (−0.38 Å) compared to X-ray data (−0.02 Å). The difference in average value between NMR and X-ray values is partly explained by a secondary peak around 0.5 Å in the X-ray distribution, which otherwise has a maximum of probability around −0.40 Å, thus close to the NMR average value. Sequence variability is important in experimental slide distributions, and the X-ray peak at 0.5 Å is mostly contributed by CG and GC steps. The slide average obtained with the AMBER parmbc0 force field in explicit solvation (−0.40 Å) is close to the NMR results in average, whereas the CHARMM all27 force field (−0.02 Å) is closer to X-ray results. However, when considering the maximum of probability, slide is lightly overestimated with the all27 force field. We have thus NMR ≈ parmbc0 < X-ray < all27. The GB method influence is consistent between the two force fields, with a general underestimation of the slide parameter, compared to explicit solvation. GBOBC1 and GBOBC2 are the most affected, then comes GBHCT, and GBMV is the closest to explicit solvation, namely, GBOBC1 ≈ GBOBC2 < GBHCT < GBMV < explicit. Distributions obtained from simulations are close to a Gaussian

Table 4. Structural Parameter Averages^a

FF	parmbc0							all27					
	solv.	NMR	X-ray	expl.	GBHCT	GBobc1	GBobc2	GBMV	expl.	GBHCT	GBobc1	GBobc2	GBMV
shift		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
slide		-0.38	-0.02	-0.40	-0.95	-1.12	-1.04	-0.52	-0.02	-0.28	-0.38	-0.40	-0.08
rise		3.33	3.34	3.34	3.48	3.51	3.42	3.35	3.36	3.50	3.61	3.53	3.39
tilt		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
roll		2.33	2.45	3.59	3.75	5.53	6.28	4.67	5.60	4.60	6.50	6.97	7.69
twist		33.87	34.10	32.44	30.96	29.65	30.25	32.70	34.15	33.50	32.09	32.44	34.40
minor		13.2	12.7	13.0	13.2	14.1	14.0	12.7	13.8	13.3	14.4	14.4	14.4
major		18.1	18.3	19.3	21.8	22.7	21.9	19.1	17.1	18.8	19.7	19.2	15.7
α		-64.7	-55.6	-73.6	-69.5	-69.9	-69.8	-68.5	-59.9	-60.1	-61.4	-61.7	-60.6
β		179.4	172.0	169.7	173.0	172.7	172.2	170.4	168.0	171.7	172.5	173.1	172.1
γ		53.5	46.2	55.6	57.4	57.5	57.4	58.3	51.7	52.3	52.0	51.8	50.2
ϵ		-177.6	-166.2	-162.1	-171.5	-172.1	-172.5	-170.5	-166.6	-169.7	-168.9	-169.8	-169.6
ζ		-99.5	-109.0	-102.3	-94.4	-90.8	-90.6	-93.1	-108.0	-106.8	-105.0	-103.4	-100.9
χ		-114.4	-110.9	-117.2	-123.6	-125.7	-124.7	-122.4	-112.3	-111.8	-111.4	-111.8	-113.9
P		145.0	145.0	133.2	125.1	120.4	122.0	126.4	143.9	146.1	145.2	146.7	146.5
τ_m		33.7	36.0	33.9	37.7	37.6	37.3	35.3	35.1	39.4	39.5	38.8	37.6

^a Structural parameters are measured with the 3DNA program on the central dimer and averaged over trajectories of the 136 sequences simulated. Global averages are obtained by applying a weighting to ensure equal composition of the 10 unique dimer steps. Angular parameters' periodicity is taken into account. Parameter averages from experimental data and explicit solvation simulations of Orozco and co-workers²⁸ are calculated following the same protocol. Averages are given in Å for translational step parameters and groove widths and in degrees for angular parameters.

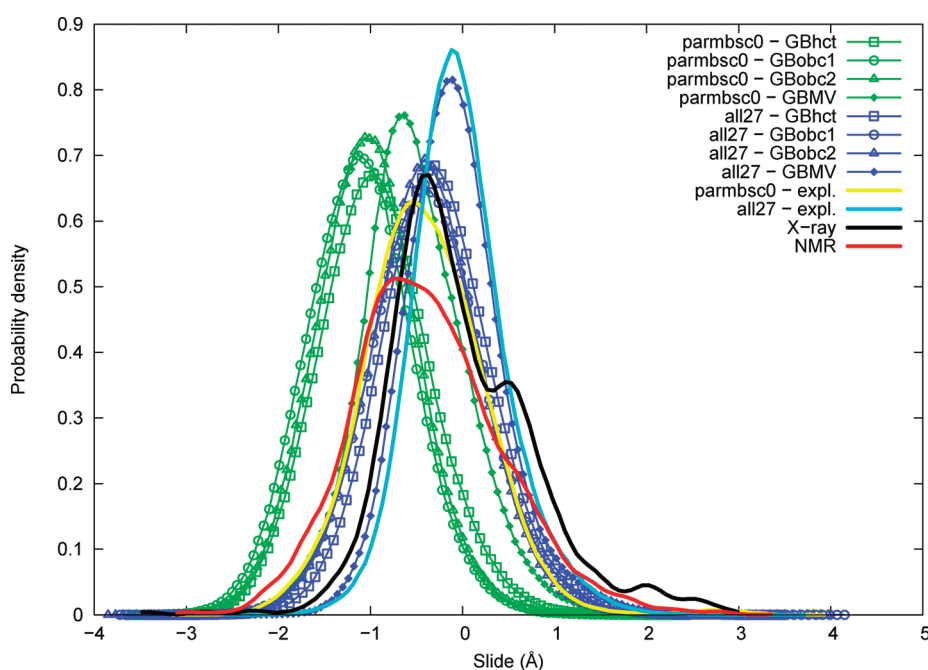


Figure 1. Slide probability distribution. Structural parameters are measured with the 3DNA program on the central dimer over trajectories of the 136 sequences simulated. Distributions are obtained from series with a kernel density estimation, using Gaussian functions as a kernel. A weighting is applied to ensure equal composition of the 10 unique dimer steps. Angular parameters' periodicity is taken into account. Parameter distributions from experimental data and explicit solvation simulations of Orozco and co-workers²⁸ are calculated following the same protocol.

shape. We see here that structural deformations resulting from the GB method choice can be of comparable magnitude to force field differences. Another point is that the GB method effect is here relatively independent of the force field.

Rise. Rise distributions are given in Figure 2. The rise average is very close between NMR (3.33 Å) and X-ray (3.34 Å). The

maximum of probability is however slightly smaller in NMR (around 3.23 Å) compared to X-ray (around 3.27 Å). Both force fields in explicit solvation are very close in average to experimental data (3.34 Å for parmbc0 and 3.36 Å for all27) but slightly higher in terms of the maximum of probability, namely, NMR < X-ray < parmbc0 ≈ all27. The AMBER GB methods

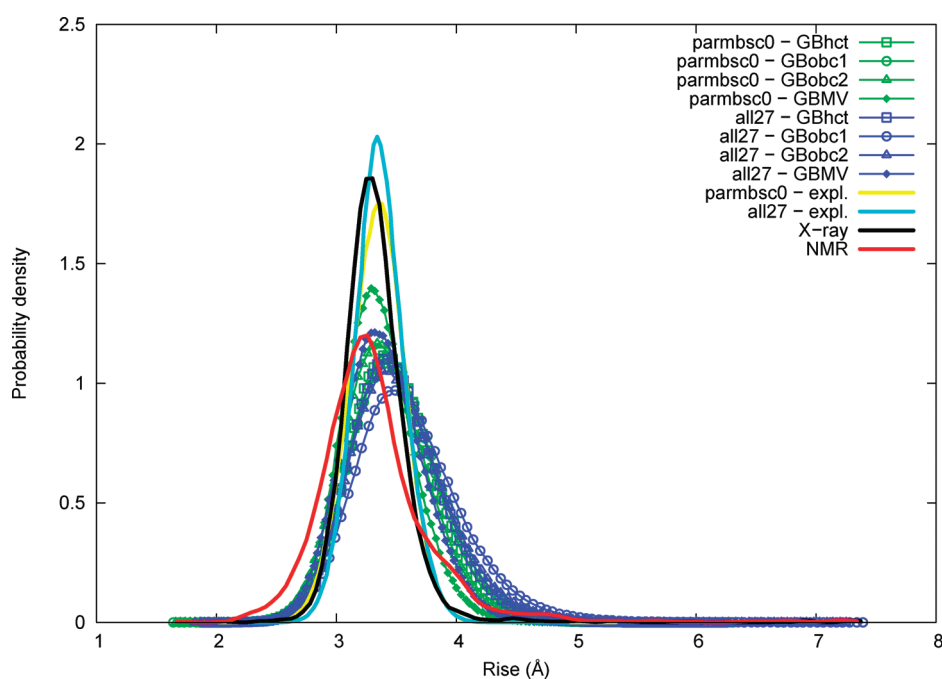


Figure 2. Rise probability distribution. See Figure 1 legend.

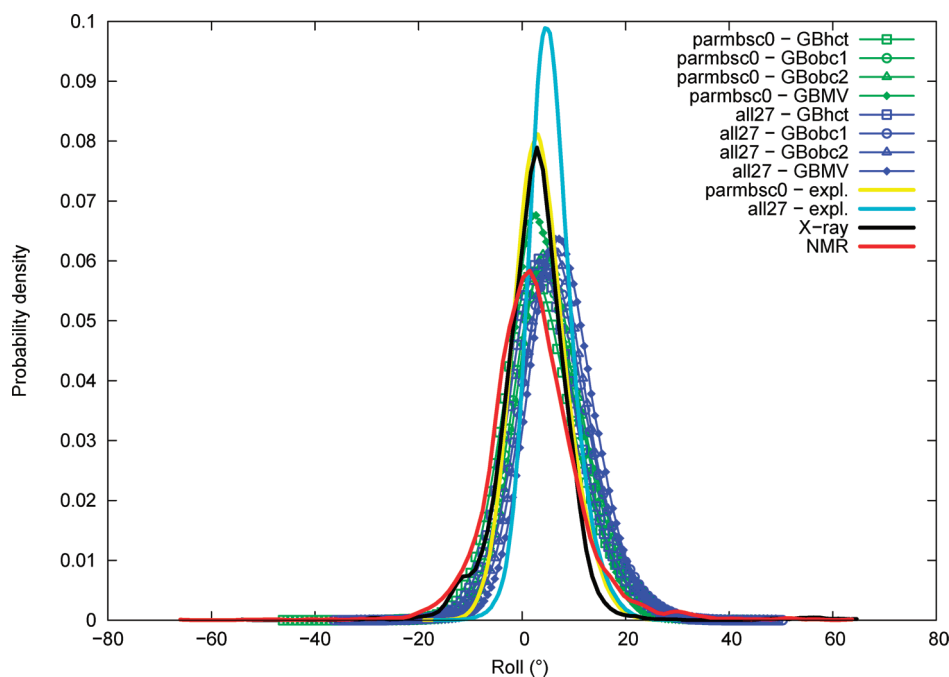


Figure 3. Roll probability distribution. See Figure 1 legend.

(GBHCT, GBOBC1, and GBOBC2) slightly overestimate rise, in particular GBOBC1, whereas the CHARMM GBMV method is close to explicit solvation, namely, explicit \approx GBMV $<$ GBHCT \approx GBOBC2 $<$ GBOBC1. Compared to slide, both the GB method and force field effects are of lower relative importance for rise.

Roll. Roll distributions are given in Figure 3. The roll average is slightly lower for NMR (2.33°) compared to that for X-ray (2.45°). Both force fields appear over-rolled, in particular the all27 force field (5.60°), while the parmbsc0 force field (3.59°) is closer to experimental values. The distributions indicate that

parmbsc0 is in very good agreement with the X-ray results, with a maximum of probability around 2.8° , we have then NMR $<$ X-ray \approx parmbsc0 $<$ all27. The influence of the GB method choice on roll is limited, and GB results stay within the range of force field or experimental data differences. We note however that among the three AMBER implementations, GBOBC1 and GBOBC2 give a slightly higher roll than GBHCT. The effect of GBMV is not clear, as it increases roll with all27 but stays in good agreement with explicit solvation when used with parmbsc0.

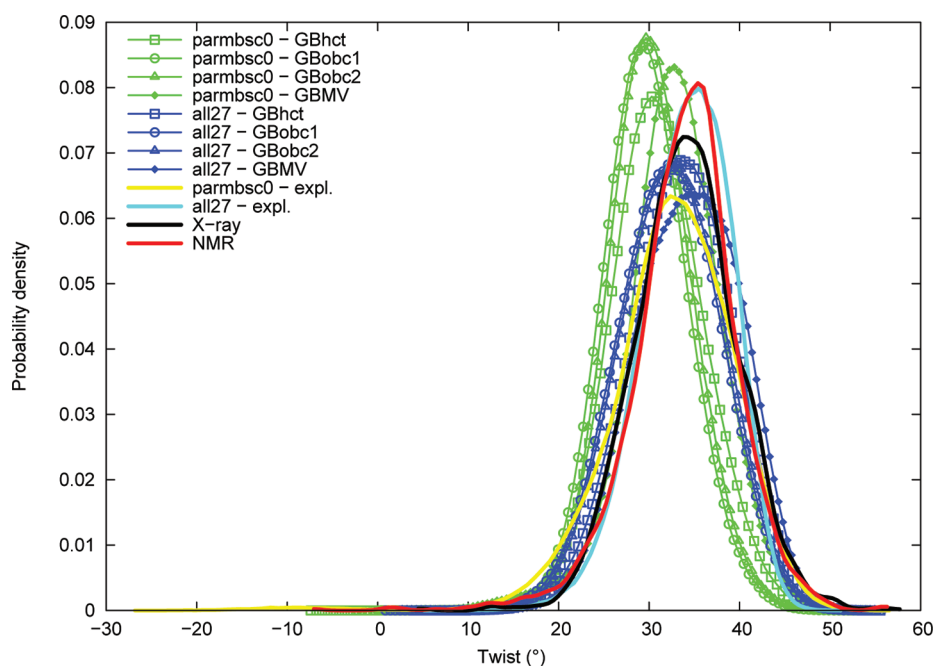


Figure 4. Twist probability distribution. See Figure 1 legend.

Twist. Twist distributions are given in Figure 4. The twist average is close between NMR (33.87°) and X-ray (34.10°). The parmbc0 force field is markedly under-twisted (32.44°), whereas the all27 force field is closer to experimental values (34.15°). The maximum of probability confirms this trend with about 35.4° for NMR, 34.1° for X-ray, 32.5° for parmbc0, and 35.4° for all27; we have thus $\text{parmbc0} < \text{X-ray} \approx \text{NMR} \approx \text{all27}$. The GB method influence on twist is consistent between the two force fields. AMBER GB methods lead to a significant twist lowering, until modal values around 29.4° for parmbc0+GBOB1, whereas GBMV is close to explicit solvation values; namely, we find $\text{GBOB1} < \text{GBOB2} < \text{GBHCT} < \text{explicit} \approx \text{GBMV}$. Both the force field and GB method effects are thus important for twist.

Groove Widths. Groove widths at the central dimer step are calculated as the average of interstrand distances P7–p4 and P8–p5 for the minor groove and as the P4–p8 distance for the major groove, where Pi is the *i*th phosphate atom of strand I counting from the 5' extremity and pj is the *j*th phosphate of strand II counting from the 3' extremity, as in El Hassan and Calladine.⁶⁹

Minor Groove. Minor groove width distributions are given in Figure 5. Experimental distributions have a broad shape and present some differences, but averages are on the same order (13.2 \AA for NMR, 12.7 \AA for X-ray), as well as maxima of probability (12.8 \AA for NMR, 13.2 \AA for X-ray). A secondary peak is present in the X-ray data at about 10.4 \AA , mostly contributed by TG, AT, and AG dimer steps. A shoulder is observed in the NMR data around 15 \AA , partly contributed by the GC step. These differences in experimental distributions are probably reflective of a lack of data at the underlying dimer level; however, convergence seems quite reasonable at the global level, as indicated by the proximity of NMR and X-ray averages and maxima of probability. The parmbc0 force field with explicit solvation lies in the experimental range (13.0 \AA), whereas the all27 force field gives a lightly wider minor groove (13.8 \AA). We have thus $\text{X-ray} \approx \text{parmbc0} \approx \text{NMR} < \text{all27}$. The GBOB1 and GBOB2 methods

tend to increase the minor groove width with both force fields. GBHCT is the closest to explicit solvation with the two force fields, in good agreement with parmbc0 and slightly lower with all27. GBMV influence is not consistent between both force fields, leading to slightly lower values than explicit solvation with parmbc0 and overestimated values with all27.

Major Groove. Major groove width distributions are given in Figure 6. Major groove width averages are on the same order between NMR (18.1 \AA) and X-ray (18.3 \AA). The maxima of probability are 17.2 \AA and 17.7 \AA for NMR and X-ray, respectively. The parmbc0 force field in explicit solvation has a noticeably overestimated major groove width (19.3 \AA); on the contrary, the all27 force field underestimates major groove width (17.1 \AA). We have thus $\text{all27} < \text{NMR} < \text{X-ray} < \text{parmbc0}$. The GBMV method lowers major groove width, in particular with the all27 force field (average value as low as 15.7 \AA). AMBER GB methods, on the other hand, tend to importantly increase major groove width in a consistent order (average value as high as 22.7 \AA with parmbc0+GBOB1); we have thus $\text{GBMV} < \text{explicit} < \text{GBHCT} < \text{GBOB2} < \text{GBOB1}$. We observe here that the major groove width is very sensible to the GB method choice, and also but to a lesser extent to the force field choice.

Backbone and Glycosidic Torsion Parameters. Backbone and glycosidic torsional parameters α , β , γ , ϵ , ζ , and χ , as canonically defined, will be discussed here. The δ torsion angle is already indirectly taken into account in sugar ring parameters. Conformational substates *gauche*–, *gauche*+, and *trans* are defined following a classical 3-fold staggered pattern as $-60 \pm 60^\circ$, $60 \pm 60^\circ$, and $180 \pm 60^\circ$, respectively.

Alpha (α). Distributions of the α backbone torsion angle are given in Figure 7. Experimental distributions of α are mostly Gaussian and slightly smaller in NMR data (average of -64.7° , maximum of probability around -66°) compared to those in X-ray (average of -55.6° , maximum of probability around -58°), both corresponding to a *gauche*– conformation. A small secondary peak around 35° is present in the X-ray distribution

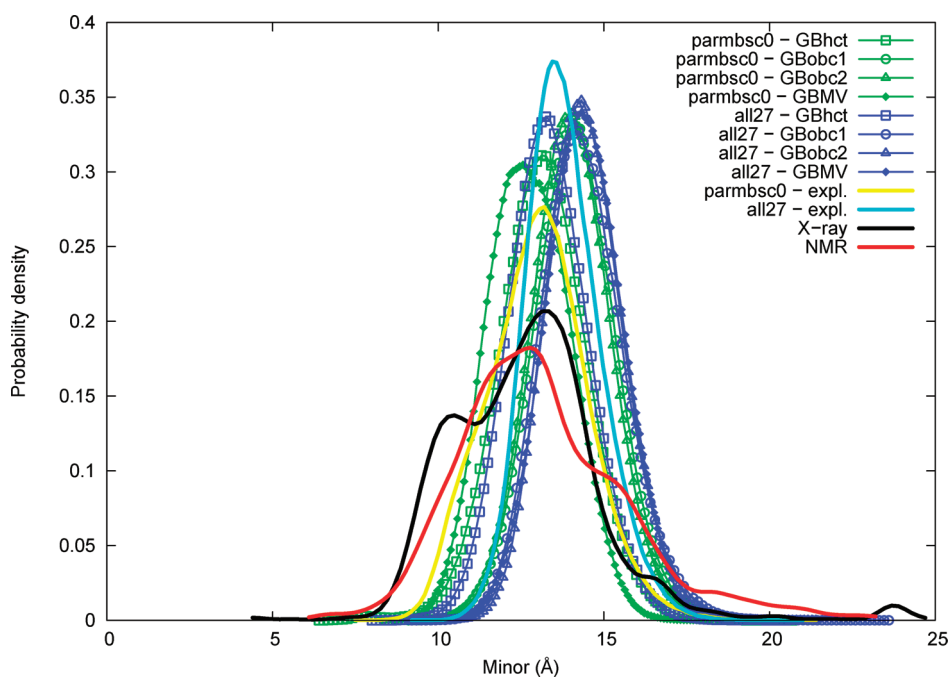


Figure 5. Minor groove width probability distribution. See Figure 1 legend.

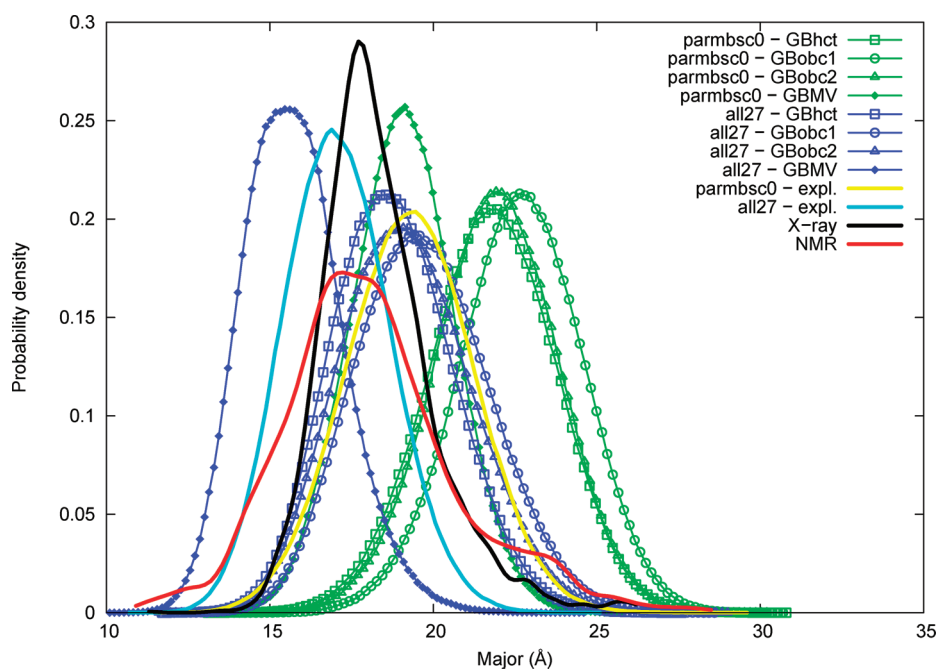


Figure 6. Major groove width probability distribution. See Figure 1 legend.

corresponding to a *gauche+* conformation. The α angle distributions from simulations are all Gaussian-shaped with no significant secondary peaks. The parmbosc0 force field in explicit solvation gives significantly lower α angles (average of -73.6° , maximum of probability around -70°), whereas the all27 force field lies in the experimental range (average of -59.9° , maximum of probability around -60°). We have thus parmbosc0 < NMR < all27 < X-ray. The use of GB implicit solvation methods leads to a slight increase of the average α torsion angle compared to explicit

solvation when used with the parmbosc0 force field and is in good agreement or leads to a slight decrease with all27. A widening of distributions is observed for both force fields when GB methods are used. We note that the GB method choice has almost no influence here, the four GB method distributions being almost superimposable.

Beta (β). Distributions of the β torsion angle are given in Figure 8. The β angle experimental distributions are mostly Gaussian centered around the *trans* conformation, with slightly

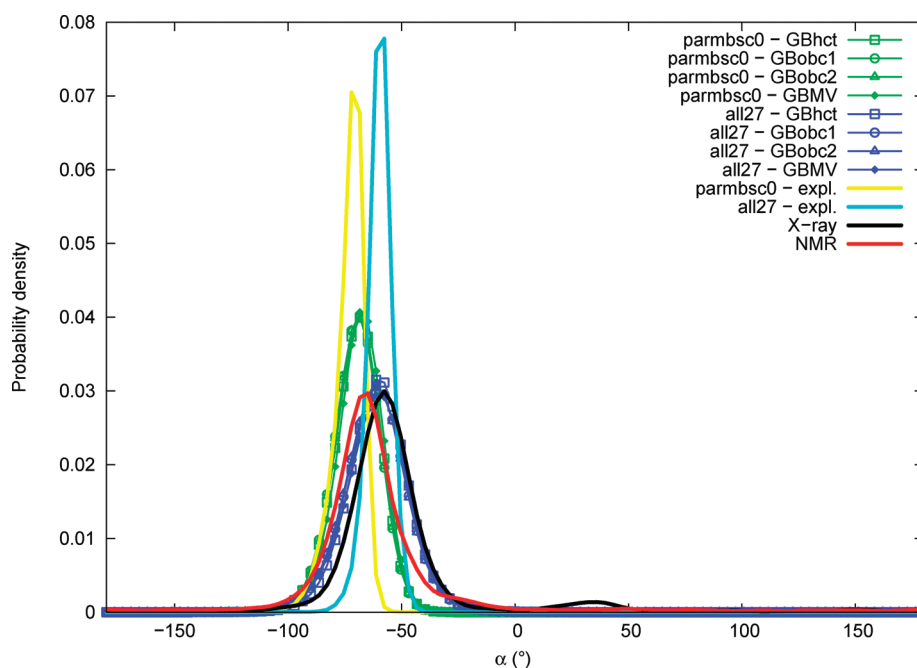


Figure 7. α torsion angle probability distribution. See Figure 1 legend.

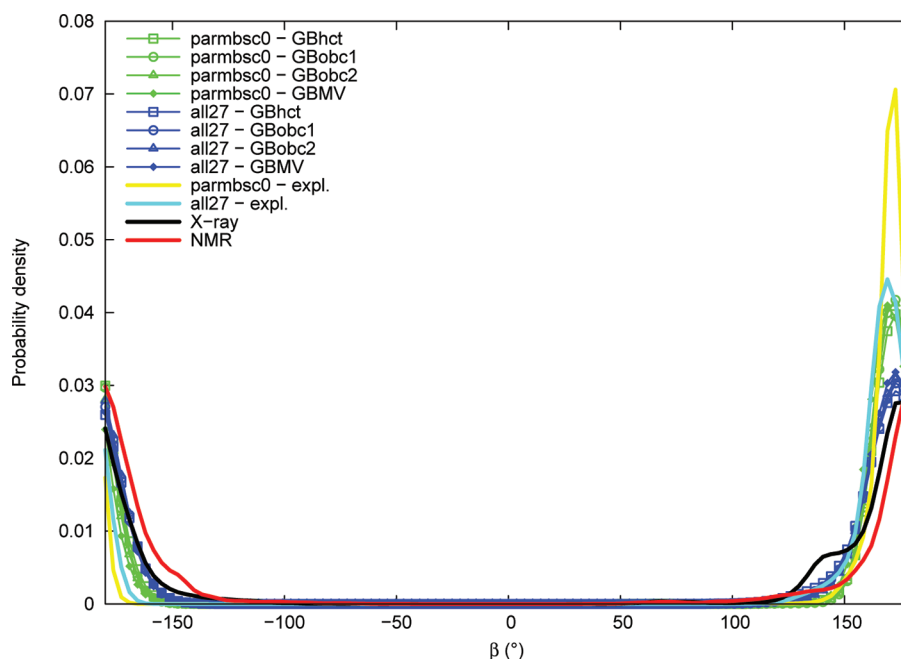


Figure 8. β torsion angle probability distribution. See Figure 1 legend.

lower values in X-ray data (average of 172.0° , maximum of probability around 174°) compared to those in NMR (average of 179.4° , maximum of probability around $\pm 180^\circ$). A small shoulder peak can be noted in the X-ray distribution around 140° . Distributions of β obtained from simulations in explicit solvation are similar to experimental ones but sharper and with slightly lower values for both all27 (average of 168.0° , maximum of probability around 169°) and parmbosc0 (average of 169.7° , maximum of probability around 171°) force fields. We have thus all27 < parmbosc0 < X-ray < NMR. The use of GB implicit solvation methods tends to increase the width of distributions

compared to explicit solvation, as well as to slightly shift the main peak toward higher values, in particular with the all27 force field. As for the α angle, differences between all GB methods are limited.

Gamma (γ). Distributions of the γ torsion angle are given in Figure 9. Experimental distributions of γ comprise a major Gaussian peak in the *gauche+* region, slightly lower in the X-ray data (average of 46.2° , maximum of probability around 47°) compared to those of the NMR (average of 53.5° , maximum of probability around 56°). Small secondary peaks can be observed for both distributions in the *trans* conformation region,

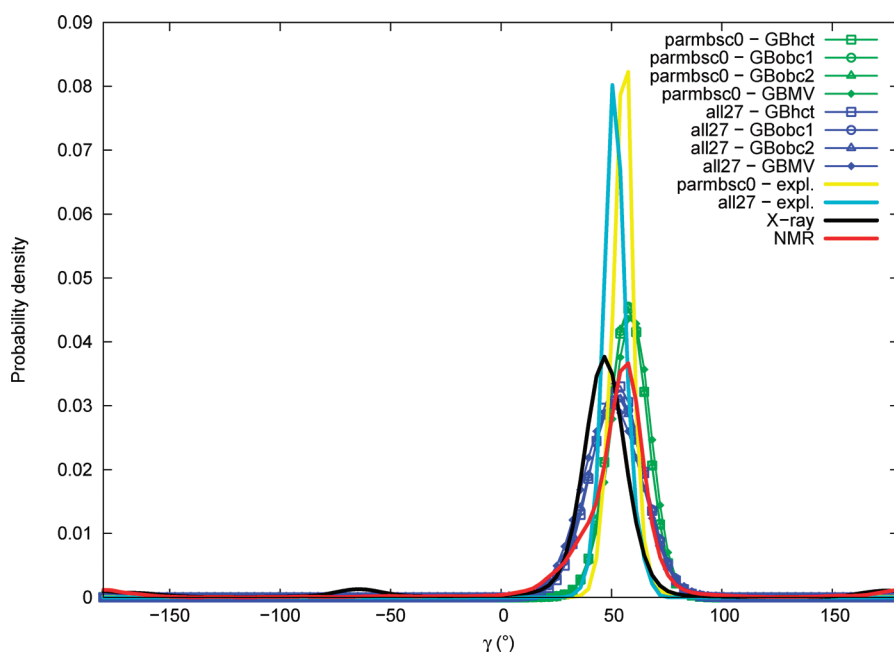


Figure 9. γ torsion angle probability distribution. See Figure 1 legend.

and in the *gauche*– region only in the X-ray distribution (around -65°). The γ angle distributions from simulations are approximately Gaussians centered in the *gauche*+ region, with no significant other peaks. Explicit solvation γ angle values are in the correct experimental range, but slightly lower with the all27 force field (average of 51.7° , maximum of probability around 52°) than with parmbsc0 (average of 55.6° , maximum of probability around 56°). We have then $X\text{-ray} < \text{all27} < \text{NMR} < \text{parmbsc0}$. GB implicit solvation methods tend to widen distributions compared to explicit solvation, as well as to slightly increase the maximum of probability peak value. Differences between GB methods are also limited here.

Epsilon (ϵ). Distributions of the ϵ torsion angle are given in Figure 10. Experimental distributions of ϵ are Gaussians centered in the *trans* conformational region, with slightly higher values in X-ray data (average of -166.2° , maximum of probability around -174°) than in NMR (average of -177.6° , maximum of probability around $\pm 180^\circ$). A small and broad secondary peak is visible in the X-ray distribution around -105° . Explicit solvation simulations also give ϵ angle distributions with a major peak in the *trans* region, close but slightly above those of X-ray with both all27 (average of -166.6° , maximum of probability about -170°) and parmbsc0 (average of -162.1° , maximum of probability about -172°). A significant secondary peak can be observed in the parmbsc0 distribution around -92° . Distributions obtained with GB implicit solvation methods are almost superimposable in the *trans* region with a maximum of probability around -172° ; small secondary peaks can be observed around -110° with all27 and around -80° with parmbsc0, with intensities depending on the GB method.

Zeta (ζ). Distributions of the ζ torsion angle are given in Figure 11. Experimental distributions of ζ comprise a major peak located in the *gauche*– region, with values slightly lower in the X-ray data (average of -109.0° , maximum of probability around -98°) compared to those of the NMR (average of -99.5° , maximum of probability around -94°). Notable secondary peaks are present in the *trans* region for both X-ray (significant peak around -174°)

and NMR (small shoulder around -170°). Distributions of ζ obtained from explicit solvation simulations also contain a major peak in the *gauche*– region, centered at values comparable to the experimental range but slightly lower with the all27 force field (average of -108.0° , maximum of probability around -103°) and slightly superior with parmbsc0 (average of -102.3° , maximum of probability around -91°). Secondary peaks in the *trans* conformational region can be observed with force fields in explicit solvation, in particular with parmbsc0 (significant peak at about 152°), and also with all27 (small shoulder around $\pm 180^\circ$). Implicit solvation ζ angle distributions are in good agreement with explicit solvation results in the *gauche*– region. Small differences between GB methods are observed concerning the precise location and intensity of the secondary peak in the *trans* region.

Chi (χ). Distributions of the χ torsion angle are given in Figure 12. Experimental distributions of χ are mostly Gaussians centered in the *gauche*– region, with slightly lower values in NMR data (average of -114.4° , maximum of probability around -114°) compared to those of the X-ray (average of -110.9° , maximum of probability around -109°). A shoulder is visible in the X-ray distribution around -153° . Explicit solvation simulations give a χ angle in the experimental range for the all27 force field (average of -112.3° , maximum of probability around -109°) and significantly underestimated with parmbsc0 (average of -117.2° , maximum of probability around -119°). The use of GB implicit solvation methods with the parmbsc0 force field further accentuates the shifting of the χ angle toward lower values by about 5° (reaching values around -125°). When GB methods are employed with the all27 force field, the χ angle main peak stays close to explicit solvation results; however, with the GBMV method, an important secondary peak appears around -151° .

Sugar Puckering Parameters. Sugar ring conformation is commonly described in terms of puckering or pseudorotation phase angle (P) and amplitude (τ_m) parameters, calculated from the five ring dihedral angles.⁷⁰ The *endo/exo* notation is sometimes used to describe particular conformations where an atom is

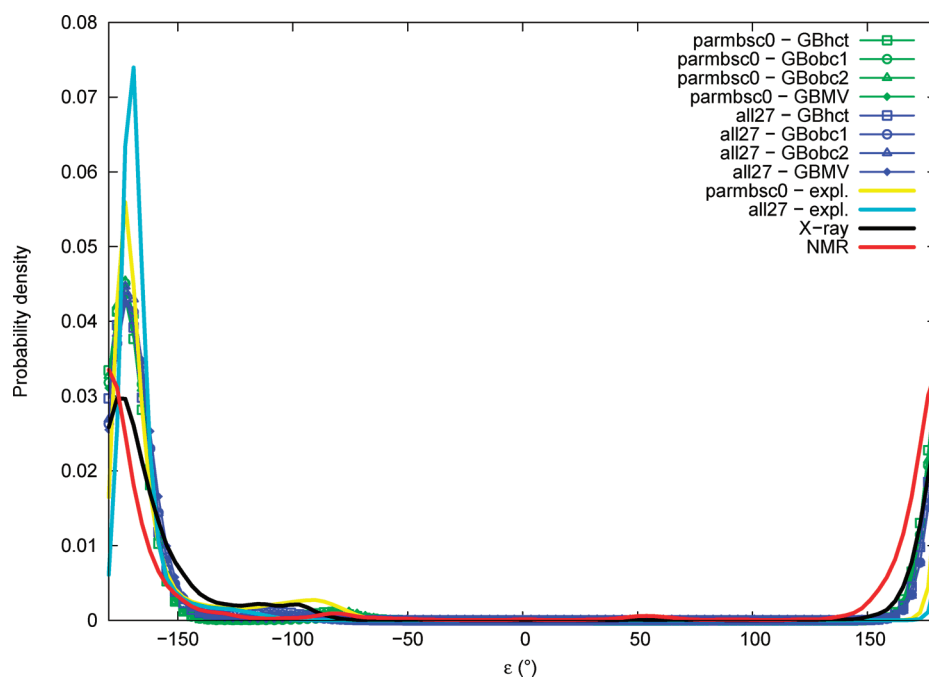


Figure 10. ϵ torsion angle probability distribution. See Figure 1 legend.

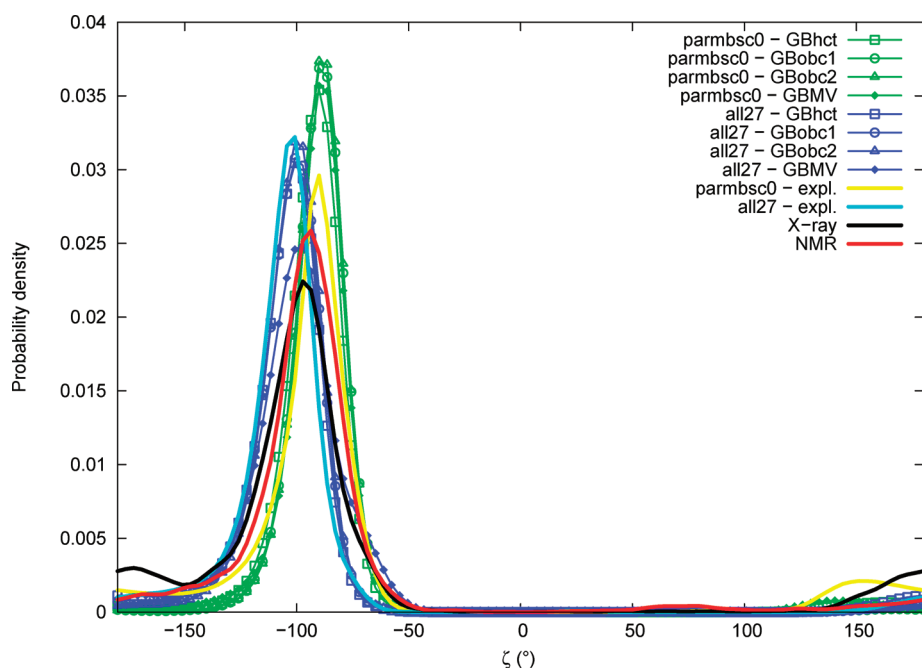


Figure 11. ζ torsion angle probability distribution. See Figure 1 legend.

out of the plane formed by the others; these conformations correspond to odd multiples of 18° of the P angle. DNA in the B form is known to favor a $C2'$ -endo ($P = 162^\circ$) deoxyribose conformation, whereas in the A form, the $C3'$ -endo ($P = 18^\circ$) conformation is preferred. Another way of referring to conformations of the sugar ring is the *North* ($P = 300-60^\circ$), *East* ($P = 60-120^\circ$), and *South* ($P = 120-220^\circ$) notation.⁷¹

Puckering Phase. Sugar puckering phase distributions are given in Figure 13. The sugar puckering phase is the same on

average in NMR and X-ray data (145.0°). Both experimental distributions present a main peak centered around 154° , corresponding to a $C2'$ -endo conformation slightly deviated toward $C1'$ -exo. Shoulders are visible around 85° ($O4'$ -endo) and 185° ($C2'$ -endo/ $C3'$ -exo) in the NMR data, and small secondary peaks can be seen in the X-ray distribution around 20° ($C3'$ -endo) and 50° ($C4'$ -exo). The parmbosc0 force field in explicit solvation has an average puckering phase markedly underestimated (133.2°), whereas the all27 force field is closer to experimental

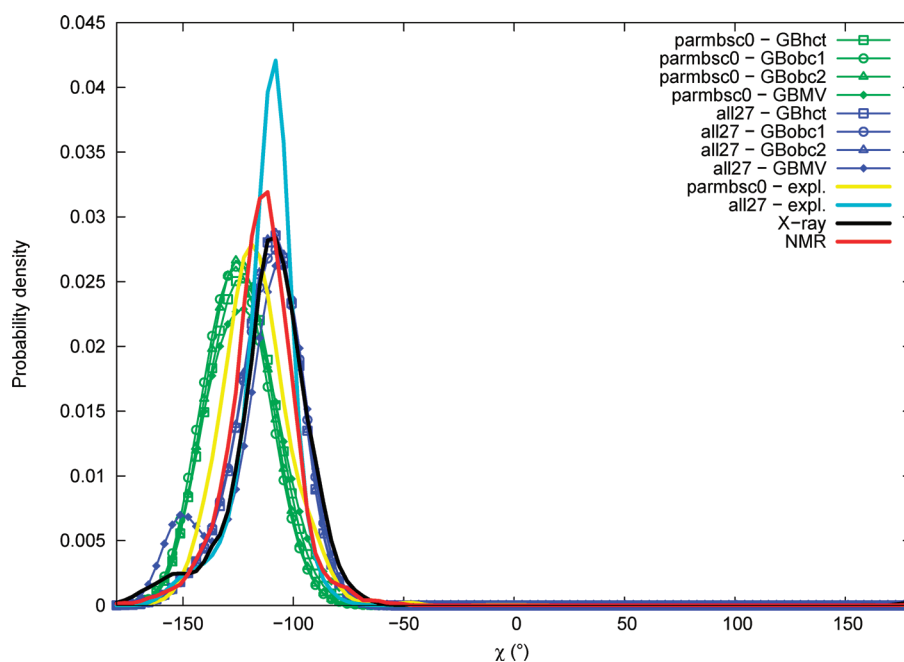


Figure 12. χ torsion angle probability distribution. See Figure 1 legend.

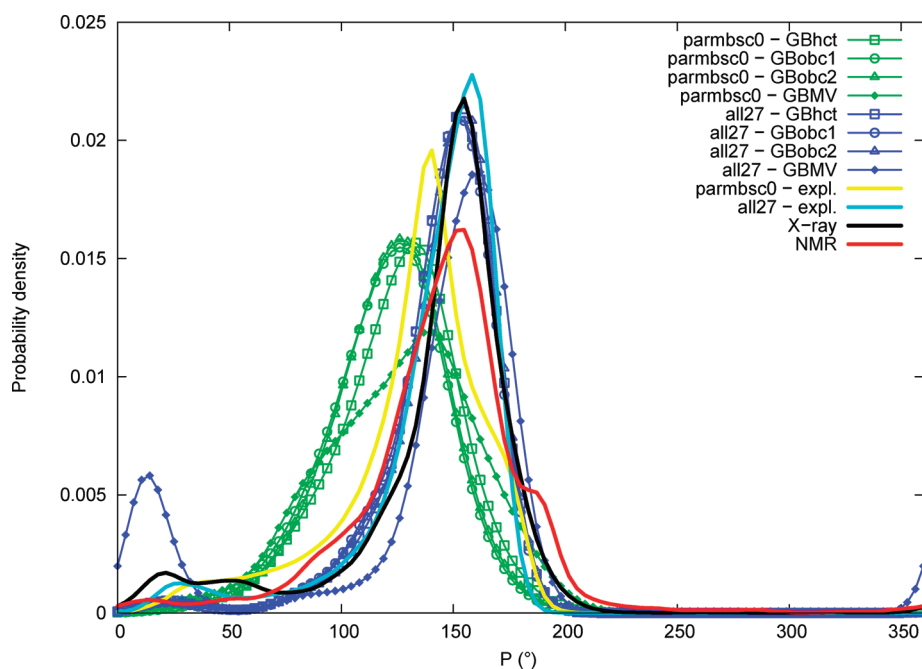


Figure 13. Sugar pucker phase (P) probability distribution. See Figure 1 legend.

data (143.9°). Distributions indicate that the main peak is indeed shifted on the left with parmbosc0 toward about 140° , indicating that the most populated conformation is closer to C1'-exo than to C2'-endo. The GB implicit solvation effect on the sugar pucker phase compared to that of explicit solvation is not straightforward. AMBER GB methods tend to lower the position of the main peak, in particular with the parmbosc0 force field, where the maximum of probability, already underestimated by parmbosc0 in itself, is further shifted by about 10° toward lower values (reaching about 125°), corresponding to a C1'-exo conformation. When using AMBER GB methods with the

all27 force field, the major peak is also shifted toward lower values but only by about 5° (at about 151°). The CHARMM GBMV method effect is quite different between the two force fields. The main peak position is retained around 140° when using GBMV with parmbosc0 but widened by an important shoulder appearing on the left in the C1'-exo region (around 115°). GBMV employed with the all27 force field slightly shifts the main peak position toward higher values (around 160°); in addition, a secondary peak around 13° becomes importantly populated. This indicates that transitions from the C2'-endo to the C3'-endo conformation (South to North) have occurred.

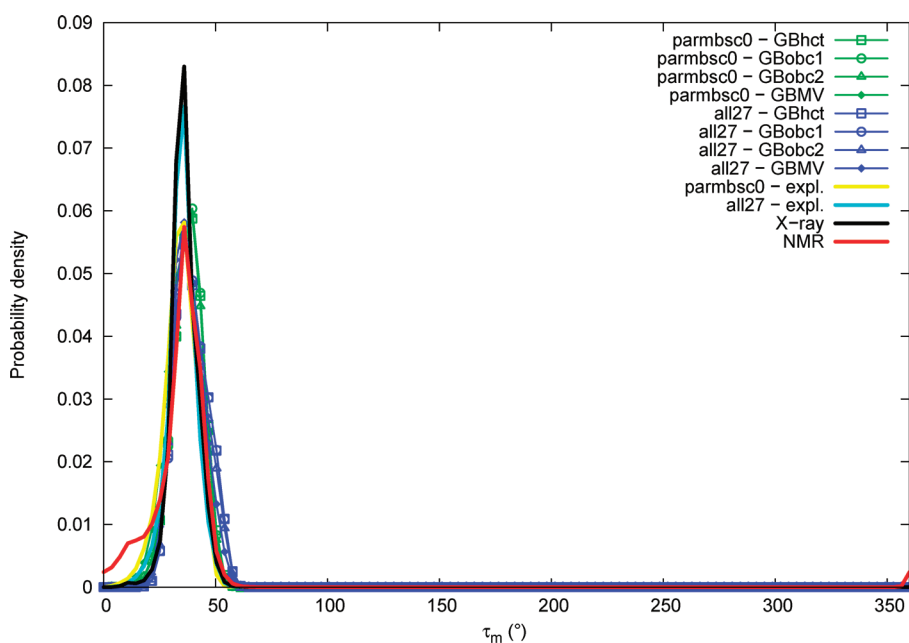


Figure 14. Sugar pucker amplitude (τ_m) probability distribution. See Figure 1 legend.

Puckering Amplitude. Sugar pucker amplitude distributions are given in Figure 14. Sugar pucker amplitude average values are on the same order between NMR (33.7°) and X-ray (36.0°). Both experimental distributions contain a major peak around 35° . The NMR distribution presents a shoulder on the left of the main peak, with a small peak around 12° . Both force fields in explicit solvation have pucker amplitudes in the experimental range (average values of 33.9° for parmbsc0 and 35.1° for all27). GB methods lead to a systematic but limited increase of the pucker amplitude average value, in a consistent order, namely, explicit < GBMV < GBOBC2 < GBOBC1 \approx GBHCT. When looking at the distributions, it appears that when GB methods are used with the parmbsc0 force field, the main peak around 35° becomes uniformly shifted until about 40° , whereas GB methods with the all27 force field retain the main peak at 35° and populate other values around 50° .

DISCUSSION

Most of the published work on DNA force field evaluation was done with explicit solvation. In this work, we attempted to evaluate DNA molecular mechanics models when employed with GB implicit solvation, using explicit solvation results and experimental data as a reference. Four GB models implemented in the AMBER suite (GBHCT, GBOBC1, GBOBC2, and GBn), and one implemented in the CHARMM program (GBMV), were tested with two different force fields, AMBER parmbsc0 and CHARMM all27. We performed 5 ns molecular dynamics simulations of 12-bp-length duplexes containing 136 different tetranucleotides, with 10 combinations of force fields and GB methods. The possibility of employing the AMBER parmbsc0 or CHARMM all27 force field with GB implicit solvation methods implemented in either the AMBER or CHARMM program was a technical requirement in this work, allowing us to examine the respective influence of force field and GB method choice on DNA simulations. Structural parameters were extracted from simulations and statistically compared to reference data. To

prevent sequence composition bias between the different data set, a weighting was applied to ensure equal contribution of the 10 different dimer steps. Considerations on the data used as a reference, experimental and explicit solvation, are first presented. The respective validity and defects of force fields and GB models are then discussed.

Experimental Data. DNA experimental structural data are mostly known from NMR spectroscopy and X-ray crystallography. It is not obvious which technique is the most reliable experimental reference for the evaluation of our simulations, as they are very different approaches. In NMR experiments, the DNA sample is in solution, as in simulations, whereas in X-ray experiments, the sample is in the solid state, which can lead to structural deformations compared to those in the aqueous phase. On the other hand, NMR data refinement importantly relies on molecular dynamics, which could bring defects from simulation models and lead to circular validation. We therefore decided to present both NMR and X-ray data. Some differences can be observed in structural parameter statistics obtained from X-ray or NMR data, concerning the precise location of the maximum of probability peak, or the detailed shape of the distribution. These differences could be the consequence of insufficient data in particular at the dimer step level but remain limited at the global level. Most of the time, conclusions are qualitatively similar when taking NMR or X-ray as a reference for the evaluation of simulation results. In addition, differences between X-ray and NMR data provide an estimation of the uncertainty expected for structural parameters.

Explicit Solvation Simulation Data. Being able to compare implicit solvation results to explicit solvation data obtained with the same force field is essential in this work, in order to assess the quality of GB models. It was necessary to find DNA explicit solvation simulation data available for the two force fields studied (parmbsc0 and all27) and obtained under the same conditions. The work of Orozco and co-workers²⁸ met these requirements and was therefore chosen as the explicit solvation reference in our study. Although the recent work of the ABC consortium²⁹ would

have provided a more extensive data set, it is only available for the parmbsc0 force field. A comparison of structural parameter averages obtained with parmbsc0 in these two systematic studies shows close agreement for most parameters. Variations in averages calculated from the same raw data between our study and those published by Orozco and co-workers are small and arise from different technical choices. It was indeed necessary to follow the same protocol and use the same program in the analysis of structural parameters between our implicit solvation results and data chosen as a reference to ensure a fair comparison.

Force Fields Evaluation. We recall here the main force field strengths and defects identified by comparing the explicit solvation simulation results of Orozco and co-workers to experimental data. Many of these conclusions have indeed already been reached in recent systematic evaluation studies of DNA simulations in explicit solvation.^{27–29} However, we believe that a summary based on our own comparison to up-to-date experimental data will prove useful here and clarify our subsequent evaluation of GB models.

The main force field defects concerning base pair step parameters are under-slide, light over-rise, light over-roll, and under-twist for parmbsc0 and over-slide, light over-rise, over-roll, and light over-twist for all27. Force field respective weaknesses are thus opposite for slide and twist and consistent for rise and roll. Mechanical couplings are known to exist between some dimer step parameters.^{72–74} For example, slide and twist are usually positively correlated, which can explain why an under- (parmbsc0) or over- (all27) estimation of one of these parameters is reflected in the other. Slide and twist are also known to be negatively correlated with roll, which is in agreement with the behavior observed with parmbsc0 but not with all27, where roll is overestimated while slide and twist are also lightly overestimated.

Opposite force field effects are found for groove widths: parmbsc0 overestimates the major groove width, whereas all27 underestimates major groove width and overestimates minor groove width. Thus, parmbsc0 increases the differences between groove widths, while all27 decreases it. Groove widths are measured over several dinucleotides; therefore, a simple and direct coupling with local dimer step parameters is not expected. However, it has been shown that increased roll values correlate with wider minor grooves and narrower major grooves.⁷⁵ The groove width deformations observed with the all27 force field could thus be linked to its main defect in terms of step parameters, that is, roll overestimation. The opposite groove width deformations observed with the parmbsc0 force field are not explained by this coupling. More complex linear relationships between slide, roll, and twist on one hand and minor and major groove widths on the other hand have been proposed by El Hassan and Calladine.⁶⁹ Applying such relations indeed leads to a small overestimation of the major groove width with parmbsc0 contributed by its under-twist defect and an underestimation with all27 contributed by its over-roll defect.

Backbone $\alpha/\beta/\gamma$ torsion parameters mostly stay in the $g-/t/g+$ conformation with both parmbsc0 and all27 force fields in explicit solvation. This conformation is canonically observed in free B-DNA,^{71,76–78} while noncanonical α/γ conformers are more abundant in protein/DNA complexes.^{71,78} Over-representation of unusual $g+/t$ conformations of α/γ angles was reported with AMBER parm94 and parm99 force fields,^{25,26,79} prompting the parmbsc0 modification. Here, we find no significant $g+/t$ conformers with all27 and less than 1% with parmbsc0. The ε and ζ torsion angles are known to adopt two main conformations,

$t/g-$ ($\varepsilon - \zeta \approx -90^\circ$) and $g-/t$ ($\varepsilon - \zeta \approx +90^\circ$), respectively designated BI and BII.^{80–82} It has been shown that the BI/BII proportion is dependent on sequence and has an important influence on helical parameters.^{83,84} The sequence averaged experimental ratio of BI/BII substates was reported to be around 80/20 for free B-DNA, from an analysis of crystal structures available in 2002,^{78,83} as well as from NMR measurements.⁸⁴ The BI/BII ratio is more in favor of BI for protein-bound DNA (around 90/10).⁷⁸ The BI/BII ratio estimated from our own analysis of experimental data is 86/14 for X-ray and 94/6 for NMR. It has to be noted that our experimental data sets contain both pure and protein bound DNA; this can be at the origin of differences in BI/BII proportions. The BI/BII ratio calculated by us from Orozco and co-workers' explicit solvation simulation data is 86/14 for parmbsc0 and 94/6 for all27. BII is thus probably under-represented with all27 and closer to the experimental proportion with parmbsc0. In a study on the Jun-Fos oligomer,⁸⁵ Hartmann and co-workers concluded that the BII state is under-represented with both parmbsc0 and all27 force fields.

The glycosidic torsion angle χ is known from earlier crystal structures surveys to adopt average values of -157° for A-DNA and -108° for B-DNA,⁸⁶ with a difference between B-DNA purines ($\chi \approx -102^\circ$) and pyrimidines ($\chi \approx -119^\circ$),⁷⁶ in conformity with our own analysis of B-DNA experimental data. The χ angle modal peak is found in the correct experimental range with the all27 force field and underestimated with parmbsc0. Deficiencies in the description of the glycosidic angle were observed by Banáš et al.⁸⁷ in simulations of RNA tetraloops with various force fields (AMBER ff94, ff99, ff99bsc0, and CHARMM all27) in explicit solvation. Reparametrization of the χ torsion of AMBER force fields was necessary to reproduce signature interactions in the tetraloops. However, it was noted that fine-tuning of the χ angle for DNA and RNA at the same time might not be possible.

In terms of the sugar pucker phase, the major conformation observed in experimental B-DNA is close to South C2'-endo. The all27 force field provides a modal conformation in the correct range, whereas the parmbsc0 force field has an incorrect East leaning behavior, leading to a different major substate still South but closer to East C1'-exo. Hartmann and co-workers⁸⁵ indeed observed that sugars attached to pyrimidines oscillate between South and East with the parmbsc0 potential, whereas sugars attached to purines with parmbsc0, as well as all types of sugars with parm98 or all27, present the correct South conformation. Sugar pucker amplitude is found in the correct experimental range for both force fields. There are known correlations between the sugar pucker phase and other parameters, such as a positive coupling with twist^{74,88,89} and with the χ torsion in the *anti* range.^{90,91} The under-twisting and χ underestimation by parmbsc0 could thus be linked to its pucker East leaning behavior, and correcting one defect could cure the others.

Generalized Born Models Evaluation. We now compare our implicit solvation results to explicit solvation simulation data and present the most important defects of GB methods. We find that, although the generalized Born model is a much simpler treatment of solvation than explicit molecules, deviations observed from explicit solvation results are often on the same order as force field differences or experimental uncertainties. Some of the conclusions presented here for GBMV may have already been qualitatively obtained by Chocholoušová and Feig;⁵⁵ however, their study focused on the Drew–Dickerson dodecamer and was

not attempting to obtain converged results in terms of DNA sequence diversity.

Concerning step parameters, the GB implicit solvation method defects are significant under-slide and under-twist with AMBER pairwise GB implementations, observed with both force fields but more pronounced with parmbsc0. Light over-rise and over-roll are also observed. Among pairwise GB methods, the GBHCT method performs better than GBOBC1 and GBOBC2. It has to be noted that pairwise GB methods implemented in AMBER cause structural deformations in the same direction as the parmbsc0 force field, thereby further accentuating its own defects. GBMV, on the contrary, is close to explicit solvation values, except for roll, where light overestimation occurs.

The minor groove width is consistently widened with GBOBC1 and GBOBC2 and correctly modeled or slightly underestimated with GBHCT. GBMV has opposite effects for the two force fields, increasing the minor groove width with all27 and decreasing it with parmbsc0. In all cases, minor groove width deformations remain close to the range of force field and experimental uncertainties. On the other hand, the influence of GB methods on the major groove width is quite spectacular in terms of deformation amplitude. The major groove width is consistently and significantly overestimated with pairwise GB methods, in particular with the parmbsc0 force field, which already inherently widens it, leading to major groove widths as high as about 22 Å with the parmbsc0+pairwise GB combinations. GBMV has an opposite narrowing effect on the major groove width, particularly pronounced with the all27 force field which already underestimates it in itself, reaching a major groove width as low as about 15 Å with the all27+GBMV combination. The best current protocols leading groove widths in the correct explicit solvation range thus appear to be parmbsc0+GBMV or all27+pairwise GB.

Backbone α , β , γ , ϵ , and ζ torsion angles are overall well modeled with GB solvation methods. Distributions are almost superimposable between the different GB implementations, and maxima of probability stay close to the explicit solvation value. There is no clear performance distinction between GBMV and pairwise GB methods, as is the case with step parameters and groove widths. One can however notice a general tendency of GB methods to slightly increase α , β , γ , and ζ angle average values and decrease ϵ . A widening of distributions can also be observed compared to explicit solvation, standard deviations becoming more comparable to experimental data, possibly reflecting improved conformational sampling of torsional angles around their canonical values with GB. The proportion of noncanonical α/γ conformers is not increased by GB methods, staying under 0.5% in all cases. The BI/BII ratio obtained with GB solvation is more in favor of BI compared to explicit solvation (around 95/5). This could be caused by insufficient sampling of the minor ϵ/ζ conformer due to limited simulation length.

GB methods present more serious defects in terms of sugar puckering angle and glycosidic torsion χ . AMBER pairwise GB methods indeed tend to decrease the puckering and χ angles main peak position when used with the parmbsc0 force field, which already underestimates the χ torsion and has an intrinsic East leaning puckering defect. Modal angle values as low as 125° for puckering and -126° for χ are reached with the parmbsc0+GBOBC combinations. When AMBER GB methods are used with the all27 force field, the puckering angle underestimation is less pronounced, and the χ angle is conformed to explicit solvation. The CHARMM GBMV method effect is quite different between the two force fields. The major peak position of

puckering and χ angles is slightly overestimated with the all27 force field and underestimated with parmbsc0 with an important shoulder appearing in the East region of the puckering angle. In addition, an important secondary peak is observed with all27 in the puckering angle North region and around -151° for the χ angle. Thus, the all27+GBMV combination appears to abnormally favor the puckering C3'-endo North conformation and a substate of χ angle values which are characteristic of the A-DNA form. Sugar puckering amplitude with GB methods is close to the correct range.

An important and perhaps unexpected result of this work is that the GB method and force field choice influences on DNA simulation quality can be of the same magnitude. As a consequence, when FF and GB defects are opposite, they can fortuitously compensate, sometimes leading to even better agreement with experimental results than explicit solvation. For example, there is defect compensation for slide, twist, and major groove width when the all27 force field is used with pairwise GB methods. Another finding is that GB influence on DNA structure is in many cases relatively independent of the force field; that is, we could observe consistent effects of GB methods on DNA structural parameters when using different force fields. For example, pairwise GB methods consistently decrease slide and twist parameters and increase major groove width with both force fields. This facilitates the analysis, as GB and FF effects can to a certain extent be considered additive. An interest of this study is to provide a quantitative estimation of structural deformations expected when simulating DNA with GB solvation methods. Overall, we can conclude that realistic DNA modeling can be obtained with GB implicit solvation methods. In particular, the GBMV method implemented in CHARMM gives the best results, comparable to explicit solvation in most cases. Defects found in AMBER pairwise GB methods are generally less pronounced or not present with the GBMV method. It has however to be noted that the GBMV method is significantly more time-consuming than pairwise GB implementations, by a factor of about 7, although this includes inherent speed differences between AMBER and CHARMM. Among AMBER GB variants, GBHCT leads to better DNA modeling than GBOBC1, GBOBC2, or GBn. In light of our results, the best combinations of force field and GB variants that we can recommend for DNA simulations are parmbsc0+GBMV and all27+GBHCT; in the latter case, compensation of defects contributes to the success of the model. Ironically, this corresponds to the GB implementation of one program with the force field generally used with the other program, that is, the "AMBARMM" and "CHAMBER" approaches. Although all27+GBMV would come next, it features worrying underestimation of the major groove, overestimation of North puckering, and glycosidic angle values characteristics of the A-DNA form. The parmbsc0+pairwise GB protocol is not recommended for DNA simulations, as in this situation, force field and GB method defects add up for slide, twist, major groove width, puckering phase, and glycosidic torsion parameters, leading to important structural deformations.

CONCLUSION

Implicit treatments of solvation like the generalized Born approach are not expected to be as detailed and realistic as the explicit simulation of solvent molecules, but they present many advantages like improved computational efficiency, enhanced conformational sampling, and the possibility of modeling large

systems and conformational changes. Implicit solvation modeling of DNA is believed to be more challenging than proteins because of the more complicated electrostatics environment. In this study, we performed a systematic evaluation of two DNA force fields with five different GB solvation models. We have shown that past decade developments have brought GB models into reasonable general agreement with explicit solvation and experimental data in terms of DNA structure and dynamics, and we identified particular defects of force fields and GB methods still present. This work provides guidance for the choice of force field and GB variant combinations in DNA simulations. The protocol developed here can be reused to evaluate force field or GB model modifications that could be suggested to improve DNA theoretical models.

AUTHOR INFORMATION

Corresponding Author

*E-mail: case@biomaps.rutgers.edu.

ACKNOWLEDGMENT

The authors wish to thank Wilma K. Olson and Guohui Zheng for valuable discussions and help in using the 3DNA program and the 3DNA Landscapes platform. This work was supported by NIH grant GM57513.

REFERENCES

- (1) Levitt, M. Computer Simulation of DNA Double-helix Dynamics. *Cold Spring Harb. Symp. Quant. Biol.* **1983**, *47*, 251–262.
- (2) Tidor, B.; Irikura, K. K.; Brooks, B. R.; Karplus, M. Dynamics of DNA oligomers. *J. Biomol. Struct. Dyn.* **1983**, *1*, 231–252.
- (3) Seibel, G. L.; Singh, U. C.; Kollman, P. A. A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 6537–6540.
- (4) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (5) Cheatham, T. E., III; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.
- (6) York, D. M.; Yang, W.; Lee, H.; Darden, T.; Pedersen, L. G. Toward the Accurate Modeling of DNA: The Importance of Long-Range Electrostatics. *J. Am. Chem. Soc.* **1995**, *117*, 5001–5002.
- (7) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (8) MacKerell, A. D., Jr.; Wiórkiewicz-Kuczera, J.; Karplus, M. An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.
- (9) Langley, D. R. Molecular dynamic simulations of environment and sequence dependent DNA conformations: the development of the BMS nucleic acid force field and comparison with experimental results. *J. Biomol. Struct. Dyn.* **1998**, *16*, 487–509.
- (10) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.
- (11) Foloppe, N.; MacKerell, A. D., Jr. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21*, 86–104.

- (12) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817–3829.
- (13) Auffinger, P.; Westhoff, E. Simulations of the molecular dynamics of nucleic acids. *Curr. Opin. Struct. Biol.* **1998**, *8*, 227–236.
- (14) Beveridge, D. L.; McConnell, K. J. Nucleic acids: theory and computer simulation, Y2K. *Curr. Opin. Struct. Biol.* **2000**, *10*, 182–196.
- (15) Cheatham, T. E., III; Kollman, P. A. Molecular Dynamics Simulation of Nucleic Acids. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435–471.
- (16) Cheatham, T. E., III; Young, M. A. Molecular Dynamics Simulation of Nucleic Acids: Successes, Limitations, and Promise. *Biopolymers* **2001**, *56*, 232–256.
- (17) Giudice, E.; Lavery, R. Simulations of Nucleic Acids and Their Complexes. *Acc. Chem. Res.* **2002**, *35*, 350–357.
- (18) Norberg, J.; Nilsson, L. Molecular Dynamics Applied to Nucleic Acids. *Acc. Chem. Res.* **2002**, *35*, 465–472.
- (19) Orozco, M.; Pérez, A.; Noya, A.; Javier Luque, F. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* **2003**, *32*, 350–364.
- (20) Cheatham, T. E., III. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360–367.
- (21) Orozco, M.; Noy, A.; Pérez, A. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.* **2008**, *18*, 185–193.
- (22) Feig, M.; Pettitt, B. M. Structural Equilibrium of DNA Represented with Different Force Fields. *Biophys. J.* **1998**, *75*, 134–149.
- (23) MacKerell, A. D., Jr.; Banavali, N. K. All-Atom Empirical Force Field for Nucleic Acids: II. Application to Molecular Dynamics Simulations of DNA and RNA in Solution. *J. Comput. Chem.* **2000**, *21*, 105–120.
- (24) Reddy, S. Y.; Leclerc, F.; Karplus, M. DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids. *Biophys. J.* **2003**, *84*, 1421–1449.
- (25) Beveridge, D. L.; Barreiro, G.; Byun, K. S.; Case, D. A.; Cheatham, T. E., III; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; et al. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.* **2004**, *87*, 3799–3813.
- (26) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* **2005**, *89*, 3721–3740.
- (27) Fujii, S.; Kono, H.; Takenaka, S.; Go, N.; Sarai, A. Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* **2007**, *35*, 6063–6074.
- (28) Perez, A.; Lankas, F.; Luque, F. J.; Orozco, M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* **2008**, *36*, 2379–2394.
- (29) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., III; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; et al. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **2010**, *38*, 299–313.
- (30) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (31) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*, 122–129.
- (32) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

- (33) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.
- (34) Jayaram, B.; Liu, Y.; Beveridge, D. L. A modification of the generalized Born theory for improved estimates of solvation energies and pK shifts. *J. Chem. Phys.* **1998**, *109*, 1465–1471.
- (35) Dominy, B. N.; Brooks, C. L., III Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.
- (36) Tsui, V.; Case, D. A. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **2001**, *56*, 275–291.
- (37) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.
- (38) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55*, 383–394.
- (39) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. Generalized Born Model with a Simple, Robust Molecular Volume Correction. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.
- (40) Schaefer, M.; Karplus, M. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Phys. Chem.* **1996**, *100*, 1578–1599.
- (41) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. Effective Atom Volumes for Implicit Solvent Models: Comparison between Voronoi Volumes and Minimum Fluctuation Volumes. *J. Comput. Chem.* **2001**, *22*, 1857–1879.
- (42) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606–10614.
- (43) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003**, *24*, 1348–1356.
- (44) Im, W.; Lee, M. S.; Brooks, C. L., III Generalized Born Model with a Simple Smoothing Function. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
- (45) Onufriev, A.; Case, D. A.; Bashford, D. Effective Born radii in the generalized Born approximation: the importance of being perfect. *J. Comput. Chem.* **2002**, *23*, 1297–1304.
- (46) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., III Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **2004**, *25*, 265–284.
- (47) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. Comparative study of generalized Born models: Protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6760–6764.
- (48) Bashford, D.; Case, D. A. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
- (49) Simonson, T. Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.
- (50) Feig, M.; Brooks, C. L., III Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
- (51) Chen, J.; Brooks, C. L., III; Khandogin, J. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.
- (52) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.
- (53) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **1999**, *101*, 426–434.
- (54) Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.
- (55) Chocholoušová, J.; Feig, M. Implicit Solvent Simulations of DNA and DNA-Protein Complexes: Agreement with Explicit Solvent vs Experiment. *J. Phys. Chem. B* **2006**, *110*, 17240–17251.
- (56) Ruscio, J. Z.; Onufriev, A. A Computational Study of Nucleosomal DNA Flexibility. *Biophys. J.* **2006**, *91*, 4121–4132.
- (57) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (58) Arnott, S.; Smith, P. J. C.; Chandrasekaran, R. In *Handbook of Biochemistry and Molecular Biology*, 3rd ed.; Fasman, G. D., Ed.; CRC Press: Cleveland, OH, 1976; Nucleic Acids, Vol. II, pp 411–422.
- (59) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (60) Crowley, M. F.; Williamson, M. J.; Walker, R. C. CHAMBER: Comprehensive Support for CHARMM Force Fields Within the AMBER Software. *Int. J. Quantum Chem.* **2009**, *109*, 3767–3772.
- (61) Chocholoušová, J.; Feig, M. Balancing an Accurate Representation of the Molecular Surface in Generalized Born Formalisms with Integrator Stability in Molecular Dynamics Simulations. *J. Comput. Chem.* **2006**, *27*, 719–729.
- (62) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.
- (63) Dickerson, R. E. Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Res.* **1989**, *17*, 1797–1803.
- (64) Lu, X. J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, S108–S121.
- (65) Lu, X. J.; Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **2008**, *3*, 1213–1227.
- (66) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (67) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **1992**, *63*, 751–759.
- (68) Zheng, G.; Colasanti, A. V.; Lu, X. J.; Olson, W. K. 3DNA Landscapes: a database for exploring the conformational features of DNA. *Nucleic Acids Res.* **2010**, *38*, D267–D274.
- (69) El Hassan, M. A.; Calladine, C. R. Two Distinct Modes of Protein-induced Bending in DNA. *J. Phys. Chem. B* **1998**, *282*, 331–343.
- (70) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. A New Description Using the Concept of Pseudorotation. *J. Am. Chem. Soc.* **1972**, *94*, 8205–8212.
- (71) Várnai, P.; Djuranovic, D.; Lavery, R.; Hartmann, B. Alpha/gamma Transitions in the B-DNA backbone. *Nucleic Acids Res.* **2002**, *30*, 5398–5406.
- (72) Gorin, A. A.; Zhurkin, V. B.; Olson, W. K. B-DNA Twisting Correlates with Base-pair Morphology. *J. Mol. Biol.* **1995**, *247*, 34–48.
- (73) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11163–11168.
- (74) Olson, W. K.; Zhurkin, V. B. Modeling DNA deformations. *Curr. Opin. Struct. Biol.* **2000**, *10*, 286–297.
- (75) Suzuki, M.; Yagi, N. An In-the-Groove View of DNA Structures in Complexes with Proteins. *J. Mol. Biol.* **1996**, *255*, 677–687.
- (76) Schneider, B.; Neidle, S.; Berman, H. M. Conformations of the Sugar-Phosphate Backbone in Helical DNA Crystal Structures. *Biopolymers* **1997**, *42*, 113–124.
- (77) Beckers, M. L. M.; Buydens, L. M. C. Multivariate Analysis of a Data Matrix Containing A-DNA and B-DNA Dinucleoside Monophosphate Steps: Multidimensional Ramachandran Plots for Nucleic Acids. *J. Comput. Chem.* **1998**, *19*, 695–715.
- (78) Djuranovic, D.; Hartmann, B. Conformational characteristics and correlations in crystal structures of nucleic acid oligonucleotides: evidence for sub-states. *J. Biomol. Struct. Dyn.* **2003**, *20*, 771–788.

- (79) Várnai, P.; Zakrzewska, K. DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.* **2004**, *32*, 4269–4280.
- (80) Gupta, G.; Bansal, M.; Sasisekharan, V. Reversible Bending and Helix Geometry in a B-DNA Dodecamer: CGCGAATT^{Br}CGCG*. *J. Biol. Chem.* **1982**, *257*, 14686–14707.
- (81) Privé, G. G.; Heinemann, U.; Chandrasegaran, S.; Kan, L.-S.; Kopka, M.; Dickerson, R. E. Helix Geometry, Hydration, and G·A Mismatch in a B-DNA Decamer. *Science* **1987**, *238*, 498–504.
- (82) Hartmann, B.; Piazzola, D.; Lavery, R. BI-BII transitions in B-DNA. *Nucleic Acids Res.* **1993**, *21*, 561–568.
- (83) Djuranovic, D.; Hartmann, B. DNA fine structure and dynamics in crystals and in solution: the impact of BI/BII backbone conformations. *Biopolymers* **2004**, *73*, 356–368.
- (84) Heddi, B.; Foloppe, N.; Bouchemal, N.; Hantz, E.; Hartmann, B. Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *J. Am. Chem. Soc.* **2006**, *128*, 9170–9177.
- (85) Heddi, B.; Foloppe, N.; Oguey, C.; Hartmann, B. Importance of Accurate DNA Structures in Solution: The Jun-Fos Model. *J. Mol. Biol.* **2008**, *382*, 956–970.
- (86) Lu, X. J.; Shakked, Z.; Olson, W. K. A-form Conformational Motifs in Ligand-bound DNA Structures. *J. Mol. Biol.* **2000**, *300*, 819–840.
- (87) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E., III; Šponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.
- (88) Zhurkin, V. B.; Lysov, Y. P.; Florentiev, V. L.; Ivanov, V. I. Torsional flexibility of B-DNA as revealed by conformational analysis. *Nucleic Acids Res.* **1982**, *10*, 1811–1830.
- (89) Kollman, P.; Keepers, J. W.; Weiner, P. Molecular-mechanics studies on d(CGCGAATTCGCG)₂ and dA₁₂·dT₁₂: An illustration of the coupling between sugar repuckering and DNA twisting. *Biopolymers* **1982**, *21*, 2345–2376.
- (90) Dickerson, R. E.; Drew, H. R.; Conner, B. N.; Wing, R. M.; Fratini, A. V.; Kopka, M. L. The Anatomy of A-, B-, and Z-DNA. *Science* **1982**, *216*, 475–485.
- (91) Gelbin, A.; Schneider, B.; Clowney, L.; Hsieh, S. H.; Olson, W. K.; Berman, H. M. Geometric Parameters in Nucleic Acids: Sugar and Phosphate Constituents. *J. Am. Chem. Soc.* **1996**, *118*, 519–529.

Enhanced Lipid Diffusion and Mixing in Accelerated Molecular Dynamics

Yi Wang,* Phineus R. L. Markwick, César Augusto F. de Oliveira, and J. Andrew McCammon

Center for Theoretical Biological Physics, Howard Hughes Medical Institute, Department of Chemistry and Biochemistry, Department of Pharmacology, University of California, San Diego, La Jolla, California 92093, United States

S Supporting Information

ABSTRACT: Accelerated molecular dynamics (aMD) is an enhanced sampling technique that expedites conformational space sampling by reducing the barriers separating various low-energy states of a system. Here, we present the first application of the aMD method on lipid membranes. Altogether, $\sim 1.5 \mu\text{s}$ simulations were performed on three systems: a pure POPC bilayer, a pure DMPC bilayer, and a mixed POPC:DMPC bilayer. Overall, the aMD simulations are found to produce significant speedup in *trans*–*gauche* isomerization and lipid lateral diffusion versus those in conventional MD (cMD) simulations. Further comparison of a 70-ns aMD run and a 300-ns cMD run of the mixed POPC:DMPC bilayer shows that the two simulations yield similar lipid mixing behaviors, with aMD generating a 2–3-fold speedup compared to cMD. Our results demonstrate that the aMD method is an efficient approach for the study of bilayer structural and dynamic properties. On the basis of simulations of the three bilayer systems, we also discuss the impact of aMD parameters on various lipid properties, which can be used as a guideline for future aMD simulations of membrane systems.

INTRODUCTION

Lipid bilayers are important components of cellular membranes. They serve as barriers against the diffusion of a large variety of biomolecules, thereby providing a structural basis of cellular compartmentalization. In recent years, molecular dynamics (MD) simulations have been increasingly used to study the dynamic properties of lipid bilayers.^{1–12} For instance, the lateral diffusion coefficient, the rotation about the lipid long axis, and the collective undulatory motions have all been investigated in MD studies.^{13–16} While these simulations have made a significant contribution to our understanding of lipid bilayers, the results are often compromised by the fact that dynamic behaviors at the lipid–water interface occur on time-scales that are not readily accessible to standard MD simulation approaches. For example, the lateral diffusion coefficient of lipids is on the order of $10^{-7} \text{ cm}^2/\text{s}$,¹⁶ which means that on average a lipid molecule travels only $\sim 10 \text{ \AA}$ in the bilayer plane during a 25-ns simulation. As a result, studying dynamic properties of complex membranes can become prohibitively expensive using atomistic MD simulations. This is exemplified by the mixing process of two or more types of lipids,^{17–24} which generates mixed bilayers that better resemble cellular membranes than bilayers composed of a single lipid species.

In this study, we investigate the application of accelerated molecular dynamics (aMD)²⁵ to enhance the lateral diffusion and mixing of lipid molecules in a bilayer. In its original form, the aMD method modifies the potential energy landscape by adding a continuous, non-negative bias potential to the energy wells below a certain threshold, while leaving the energy barriers above this threshold unaffected.²⁵ As a result, barriers separating adjacent energy basins are reduced, allowing the system to sample conformational space with greater efficiency. Two parameters control the “amount” of acceleration introduced by aMD:

the acceleration threshold energy E , which determines the portion of the energy surface affected by aMD, and the acceleration factor α , which determines how smooth the modified potential surface becomes.²⁵ The aMD method has been successfully applied in multiple studies of protein systems, such as GB3,²⁶ Ubiquitin,²⁷ Ras,²⁸ and the maltose binding protein.²⁹

Here, we report the first application of aMD on lipid membranes, focusing on the effect of different aMD parameters on the *trans*–*gauche* isomerization, lateral diffusion, and mixing of a bilayer system. Simulations totaling $\sim 1.5 \mu\text{s}$ were performed on three systems: a pure POPC bilayer, a pure DMPC bilayer, and a mixed POPC:DMPC bilayer. The two lipid species allow us to compare the effect of aMD on lipids with distinct properties: POPC has a saturated palmitoyl (16:0) chain and an unsaturated oleoyl chain (18:1), while DMPC has the saturated myristoyl (14:0) chain in both tails. We demonstrate that aMD can significantly enhance the *trans*–*gauche* isomerization and lateral diffusion for both lipid species. Furthermore, comparison with conventional MD (cMD) simulations reveals a significant speedup in lipid mixing in the aMD simulation of the POPC:DMPC bilayer. In the remainder of the text, we will first explain the design of the simulations and then discuss the results of the three systems mentioned above. Our results demonstrate that the aMD method is an efficient approach to studying lipid bilayers, which can also be extended to systems involving membrane proteins. The parameters examined in our simulations provide the first set of benchmarks for lipid aMD simulations, which may be used as a guideline for parameter selection in future aMD studies of membrane systems.

Received: June 22, 2011

Published: August 24, 2011

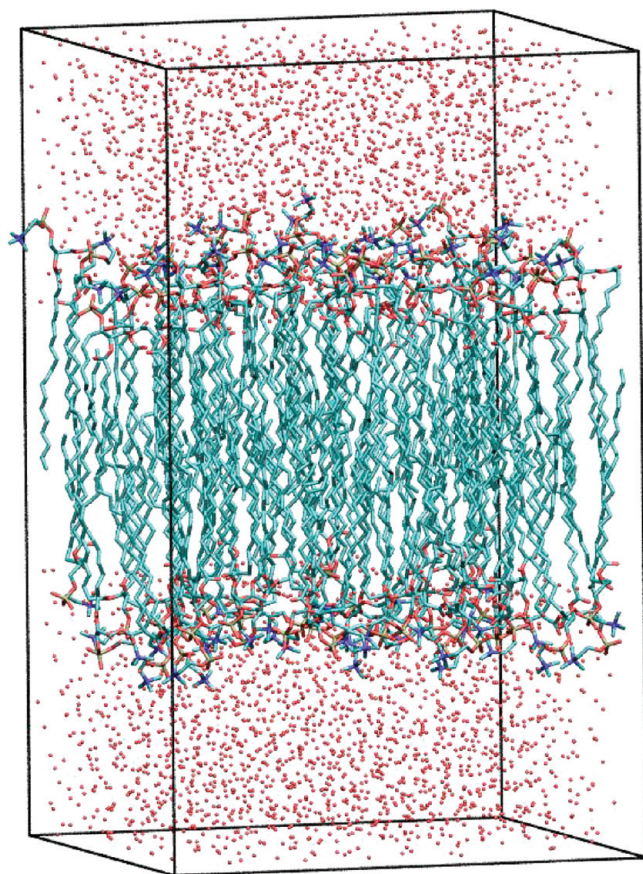


Figure 1. The initial structure of the POPC bilayer simulation system.

METHODS

System Preparation. A POPC bilayer was initially constructed using the *membrane* plugin of VMD³⁰ with 39 molecules in each leaflet. The membrane normal was placed along the z axis, and a 15-Å water layer was added to each side of the bilayer. The final system consists of 78 POPC molecules and 3693 water molecules. Since this study involves a comparative analysis of *trans*–*gauche* isomerization, the initial structure of the bilayer was designed to have a primarily *trans* conformation in the lipid tail region (Figure 1). However, simulating such a system directly under constant temperature and pressure (NPT) conditions results in erroneous artifacts, including significant interdigitation of the lipid tails and a considerable decrease in the bilayer thickness. Therefore, two test simulations were performed first, namely, a 1-ns simulation under constant temperature and volume (NVT) conditions, during which only the lipid tails were free to move, followed by a 10-ns NPAT (constant temperature, pressure, and surface area) simulation. The average volume of the system during the last nanosecond of the NPAT simulation was calculated and considered the target volume. Then, the following equilibration protocol was used to prepare the system for the production runs: Starting from the POPC system built initially (Figure 1), we performed a ~ 10 ps NPT simulation with the phosphorus atoms constrained at their initial positions using a spring constant of 5 kcal/mol/Å², which was followed by a 20 ps NPAT run. These two short

Table 1. cMD and aMD Simulations Performed for the POPC, DMPC, and Mixed POPC:DMPC Bilayers^a

lipid	simulation	dE	$d\alpha$	t (ns)
POPC	cMD1			70
	cMD2			70
	E1	10	3	70
	E2a1	30	1	10
	E2a2	30	3	10
	E2a3	30	30	70
	E3a1	40	10	10
	E3a2	40	20	70
	E3a3	40	40	70
DMPC	cMD1			70
	cMD2			70
	E1	8.8	2.6	70
	E2a1	26.4	0.88	70
	E2a2	26.4	2.6	70
	E2a3	26.4	26.4	70
	E3a1	35.2	8.8	70
	E3a2	35.2	17.6	70
	E3a3	35.2	35.2	70
mix	cMD			300
	aMD	37.6	18.8	70

^aThe unit of aMD parameters (dE and $d\alpha$) is kcal/mol/lipid.

simulations brought the system to the target volume, while allowing most of the lipid tails to maintain their *trans* conformation. The final system has an area per lipid of 68.5 Å², which is similar to the experimentally determined value for POPC in the L_{α} phase (68.3 ± 1.5 Å²).³¹

A DMPC bilayer was constructed using the above POPC bilayer as a template, which contained 78 DMPC molecules and 3535 water molecules. Following a protocol similar to that described above, this system was brought to the target volume with an area per lipid of 60.3 Å², comparable to the two reported experimental values for DMPC: 59.7 Å²³² and 60.6 Å².³³ Additionally, a mixed bilayer was generated by combining half of the DMPC with half of the POPC system. The mixed bilayer contains 39 DMPC, 39 POPC, and 3532 water molecules, with an area per lipid of 64.9 Å².

aMD: The method. In the original form of aMD,²⁵ when the potential energy of the system falls below a threshold energy, E , a boost potential is added, such that the modified potential, $V^*(\mathbf{r})$, is related to the original potential, $V(\mathbf{r})$, via

$$V^*(\mathbf{r}) = V(\mathbf{r}) + \Delta V(\mathbf{r}) \quad (1)$$

where $\Delta V(\mathbf{r})$ is the boost potential

$$\Delta V(\mathbf{r}) = \begin{cases} 0 & V(\mathbf{r}) \geq E \\ \frac{(E - V(\mathbf{r}))^2}{\alpha + E - V(\mathbf{r})} & V(\mathbf{r}) < E \end{cases} \quad (2)$$

In the above equation, E is the threshold energy specified by the user, which controls the portion of the potential surface affected by the bias, and α is the acceleration factor that determines the shape of the modified potential: when E is fixed, the smaller α is, the more flattened the energy surface becomes.

From an aMD simulation, the ensemble average of an observable, $\langle A \rangle$, can be calculated via the following reweighting

equation:

$$\langle A \rangle = \frac{\langle A(\mathbf{r}) \exp(\beta \Delta V(\mathbf{r})) \rangle^*}{\langle \exp(\beta \Delta V(\mathbf{r})) \rangle^*} \quad (3)$$

in which $\beta = 1/k_{\text{B}}T$ and $\langle \dots \rangle^*$ represents the ensemble average in the aMD ensemble.

aMD Simulation Parameters. Following the preparatory simulations described earlier, two cMD and seven aMD simulations were performed under NVT conditions for each of the POPC and DMPC bilayers (Table 1). All of the aMD simulations listed in Table 1 were performed using the “boosting dihedral” mode (aMDd).^{25,34} Additionally, a dual-boost aMD (aMDdual)^{35,36} simulation was performed for the POPC bilayer, which will be described in more detail in the Results and Discussion section. Similar to previous aMD studies,^{27,37,35} the average dihedral energy of the system during a cMD simulation was used as a reference to set the aMD parameters E and α , *i.e.*, $E = V_{\text{avg}} + dE * N$ and $\alpha = d\alpha * N$, where V_{avg} is the average dihedral energy of the system in the first 1 ns of the cMD simulation, $N = 78$ is the number of lipid molecules, and dE and $d\alpha$ are constants with units of kcal/mol/lipid. For the pure POPC bilayer, three acceleration threshold energy values (E1, E2, and E3) were examined, with $dE = 10, 30$, and 40 kcal/mol/lipid, respectively. For the last two acceleration thresholds, three independent simulations were performed with different α values (a1, a2, and a3). Note that while E1, E2, and E3 uniquely identify the threshold energy used in an aMD simulation, a1, a2, and a3 only distinguish simulations within the same threshold energy level, *i.e.*, E2a2 and E3a2 have different α values (see Table 1).

In order to keep the acceleration level comparable for the POPC and DMPC bilayers, the aMD parameters for the pure DMPC bilayer were set using the same equations described above, with dE and $d\alpha$ scaled by $\text{dihe}_{\text{DMPC}}/\text{dihe}_{\text{POPC}}$, *i.e.*, the ratio of the number of dihedrals in the DMPC and POPC bilayers.

With the exception of three POPC aMD runs, all cMD and aMD simulations mentioned above were performed for 70 ns. Three POPC aMD simulations, E2a1, E2a2, and E3a1, revealed artificial *cis*–*trans* conformational changes and were terminated at $t = 10$ ns. These simulations are discussed in more detail in the Results and Discussion section. Following each 70-ns aMD run, a 10-ns cMD simulation was performed, the starting structure of which was taken from the snapshot with the highest boost potential in the last 5 ns of the aMD trajectory. We will refer to them as “-eq” trajectories of the corresponding aMD runs; *e.g.*, E1-eq refers to the 10-ns cMD simulation following the aMD run E1.

For the mixed POPC:DMPC bilayer, one cMD and one aMD simulation were performed. The aMD parameters were chosen on the basis of the E3a2 simulations of POPC and DMPC (see Table 1). The aMD simulation of the mixed bilayer was performed for 70 ns, while the cMD simulation was performed for 300 ns to allow sufficient lipid mixing events to be sampled.

Simulation Protocols. As this study was initiated before the latest CHARMM36 force field³⁸ was available, the CHARMM27r force field for lipids was used.^{39,40} The cMD simulations were performed with the 2.7b1 release of NAMD,⁴¹ while the aMD simulations were performed with the recent NAMD implementation of aMD,³⁴ now available in the 2.8 release of the software. All simulations were performed using a time step of 2 fs, with bonds involving hydrogen atoms constrained using RATTLE⁴² and water

geometries maintained using SETTLE.⁴³ The multiple-time-stepping algorithm was used, with short-range forces calculated every step and long-range electrostatics calculated every 2 steps. The cutoff for short-range nonbonded interactions was set to 12 Å, with a switching distance of 10 Å. Assuming periodic boundary conditions, the particle mesh Ewald (PME) method⁴⁴ with a grid density of at least $1/\text{Å}^3$ was employed for computation of long-range electrostatic forces. The temperature was maintained at 303 K for all simulations using Langevin dynamics with a damping coefficient of 1 ps^{-1} . In NPT or NPAT simulations, the pressure was kept constant at 1 atm using a Nosé–Hoover–Langevin piston.⁴⁵

RESULTS AND DISCUSSION

Performance of aMD. In line with a previous study,³⁴ the aMD simulations performed with NAMD are slightly slower than the corresponding cMD runs. Using a local cluster with InfiniBand connections, the cMD and aMD simulations gave on average 0.0758 days/ns and 0.0782 days/ns on 36 processors or 0.0624 days/ns and 0.0680 days/ns on 48 processors, respectively. The 3–9% slowdown in aMD is primarily due to an extra round of energy reduction calls.³⁴ More comprehensive evaluation of the aMD performance can be found in the aforementioned study.

Pure POPC and DMPC Bilayers. *trans/gauche* Ratio. For both POPC and DMPC bilayers, the probability distributions of the lipid tail dihedrals were constructed using the last 20 ns of the 70-ns cMD simulations. Since the *trans*–*gauche* isomerization is a rapid process, the above results can be viewed as equilibrium distribution profiles. As shown in Figure S1 (Supporting Information), most dihedrals exhibit similar distributions, with three notable exceptions in the POPC bilayer: C7–C8–C9–C10, C8–C9–C10–C11, and C9–C10–C11–C12, which all involve the *cis* double bond in the unsaturated oleoyl chain. On the basis of the profiles in Figure S1, we calculated the *trans/gauche* ratio with the *trans* conformation defined as $\chi \leq -150^\circ$ or $\chi > 150^\circ$ and the *gauche* conformation defined as $-97^\circ < \chi \leq -37^\circ$ or $37^\circ < \chi \leq 97^\circ$. Due to their unique distribution profiles, the three dihedrals mentioned above are not included in the calculation of the *trans/gauche* ratio for POPC.

As shown in Figure S2 (Supporting Information), while the simulations started with a primarily *trans* conformation, the *trans/gauche* ratio became stabilized shortly after the simulations began. This fast isomerization is consistent with results from earlier studies.⁴⁶ To quantify the difference between cMD and aMD results, we calculated the equilibrium *trans/gauche* ratio using the last 20 ns of each cMD and aMD run and recorded the time when a simulation first reached its equilibrium value. Overall, the isomerization appears to be slower in POPC than DMPC: The equilibrium *trans/gauche* ratio was reached in 526 and 727 ps in the two cMD simulations of POPC, while the corresponding numbers are 55 and 82 ps in the two cMD simulations of DMPC. For both lipid species, aMD produced a significant speedup, with the fastest isomerization occurring within 81 ps for POPC and 13 ps for DMPC, corresponding to a 7.7- and 5.3-fold speedup from the cMD simulations, respectively. The slower isomerization in POPC may be related to its longer hydrocarbon tails and a lower equilibrium *trans/gauche* ratio (2.08) than DMPC (2.44). It should be noted that since the simulations started with the area per lipid corresponding to an L_α -phase bilayer, these results only reflect the difference

of cMD and aMD in promoting *trans*–*gauche* isomerization and should not be compared to a gel-to-fluid phase transition.

Lateral Diffusion Coefficient. The lipid lateral diffusion coefficient D was calculated according to the Einstein relation:

$$D = \frac{1}{N} \sum_i \frac{1}{4t} \langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \rangle \quad (4)$$

where N is the number of lipids in the system, $\mathbf{r}_i(t)$ and $\mathbf{r}_i(0)$ are the coordinates of the center-of-mass (COM) of a lipid molecule in the membrane plane at time t and time 0, respectively. To allow for sufficient equilibration, the last 60 ns of each cMD and aMD simulation were used to calculate D . The time origin $t = 0$ was shifted along the simulation trajectory to make use of all data points, and the mean square displacement (MSD) data from $t = 1$ ns to $t = 20$ ns were used to determine D by a least-squares method.

The lateral diffusion coefficients obtained from the above calculations are listed in Table 2. To quantify the comparison of cMD and aMD results, we took the average of D obtained from the two cMD simulations of each bilayer and used it to calculate the relative lateral diffusion coefficients (D_{relative}) for each

Table 2. Lateral Diffusion Coefficients of POPC and DMPC Calculated from cMD and aMD Simulations^a

simulation	POPC		DMPC	
	D (10^{-8} cm ² /s)	D_{relative}	D (10^{-8} cm ² /s)	D_{relative}
cMD1	6.2	98%	6.4	107%
cMD2	6.4	102%	5.5	93%
E1	10.4	164%	8.9	149%
E2a1			17.2	287%
E2a2			12.6	212%
E2a3	15.6	247%	9.1	152%
E3a1			17.0	285%
E3a2	15.2	240%	13.9	233%
E3a3	10.8	172%	10.8	181%

^a D_{relative} is defined as $2D/(D1 + D2)$, where $D1$ and $D2$ are the diffusion coefficients of the cMD1 and cMD2 simulations, respectively.

simulation. As shown in Figure 2 and Table 2, significant speedup is observed in all aMD simulations, with a maximum speedup of 247% for POPC and 287% for DMPC in the aMDd runs. Furthermore, a 349% speedup is observed in the aMDdual simulation of POPC. The difference in the two lipid species and the effect of aMD parameters on lipid lateral diffusion are discussed in the next section.

Recently, Roark and Feller¹⁶ showed that for a small system, the correlation length of monolayer COM motions was comparable with the dimension of the simulation unit cell. As a result, a small system may produce an artificially large D .⁷ Removing the monolayer COM prior to the calculation alleviates this problem, although such a treatment also introduces another artifact,⁴⁷ resulting in the underestimation of the lateral diffusion coefficient.¹⁶ In this study, the monolayer COM motions are removed before the calculation of D , because it produces the highest statistical precision for our simulation data (see Figure S3, Supporting Information). While we note that this treatment still suffers from the finite size effect described by Yeh and Hummer,⁴⁷ such an effect is present in both cMD and aMD simulations and, therefore, should not affect the comparison of the two methods significantly.

Effect of E and α . Further analysis of the lateral diffusion coefficients allows us to examine the effect of aMD parameters: Overall, larger values of E and smaller values of α produce faster lateral diffusion, which is explained by a higher boost potential as a result of the changes in E and α . For instance, with dE kept at 40 kcal/mol/lipid, decreasing $d\alpha$ from 40 (E3a3) to 20 (E3a2) kcal/mol/lipid increased the POPC lateral diffusion coefficient by 40%. Meanwhile, with $d\alpha$ kept at 2.6 kcal/mol/lipid, increasing dE from 8.8 (E1) to 26.4 (E2a2) kcal/mol/lipid for DMPC increased D by 42%. Such a general relationship between the acceleration level and the two aMD parameters has been well established in earlier aMD studies of peptides or proteins.^{25,26,34,48}

Comparison of the DMPC E1 and all of the E3 simulations suggests that the parameter E plays the dominant role in these aMD runs, since the increase in E determines the speedup in lipid lateral diffusion. Similar observations can be made from the comparison of E1 and E2a3 simulations of POPC. These results

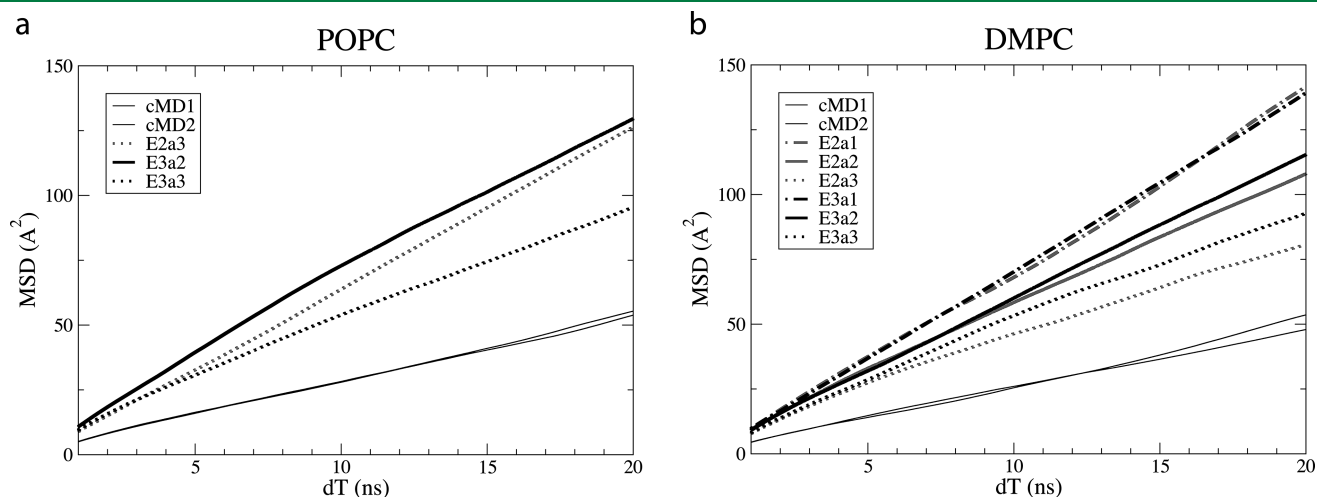


Figure 2. The mean square displacement (MSD) of POPC (a) and DMPC (b) calculated from cMD and aMD simulations. The cMD results are shown in thin black lines, while the aMD results are shown in thick lines. Results from different aMD simulations are distinguished by their colors and line styles. The corresponding lateral diffusion coefficients are listed in Table 2.

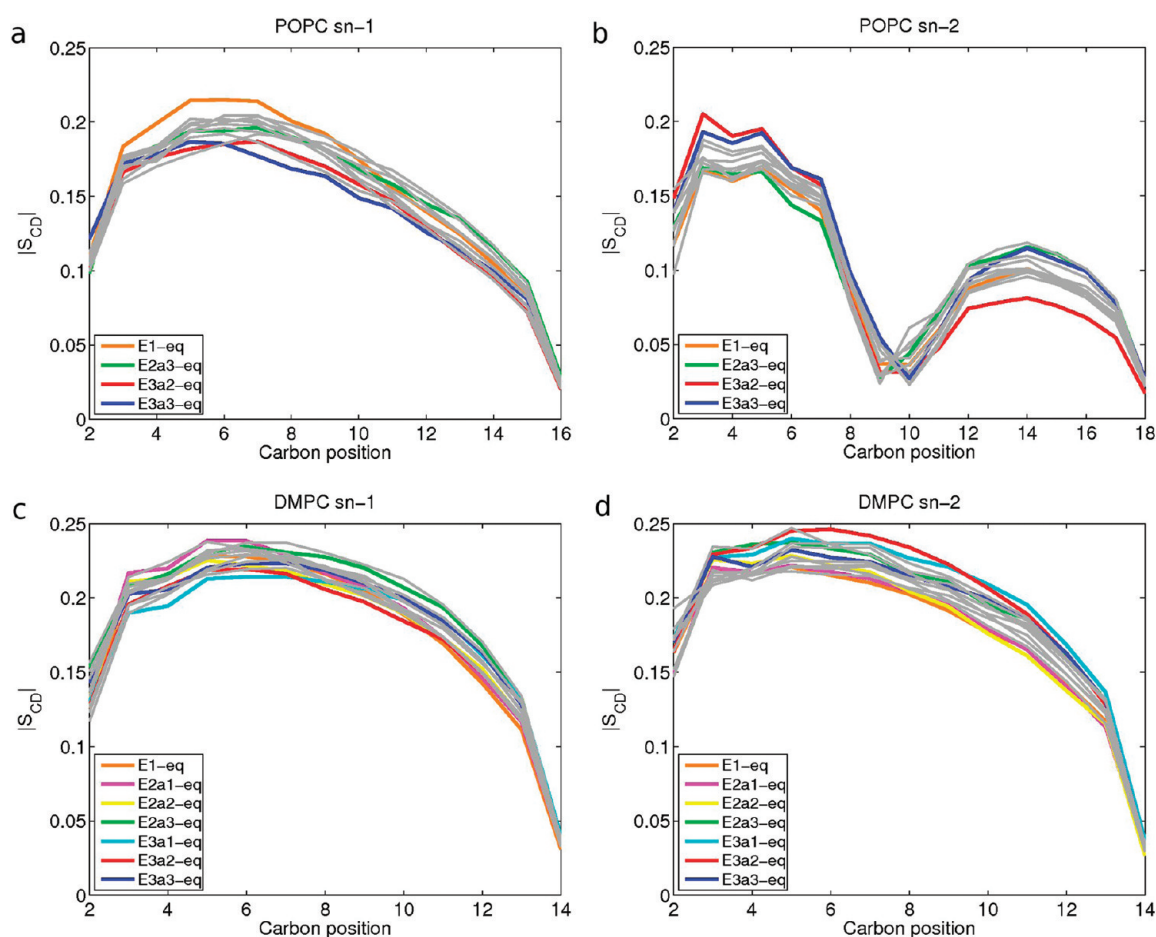


Figure 3. Order parameter S_{CD} of POPC (a, b) and DMPC (c, d) in aMD-eq simulations. Calculations were performed using the last 5 ns of the 10-ns aMD-eq simulations (colored lines). For comparison, the same calculation was repeated using a 5-ns block for the last 20 ns of the cMD1 and cMD2 simulations (eight gray lines).

suggest that when the initial acceleration threshold energy is low, increasing E is very efficient in raising the acceleration level. However, once E becomes large, α often assumes the dominant role in determining the speedup of the lateral diffusion. For example, when dE is reduced from 40 (E3a3) to 30 (E2a3) kcal/mol/lipid, decreasing α still results in larger D for POPC, suggesting that the impact of α outweighs that of E in these cases.

The analysis of all aMD simulations listed in Table 1 reveals that POPC is more sensitive to the change in α than DMPC. This difference may be explained by the different structures of the two lipids: the double bond in the sn-2 oleoyl chain of POPC produces a large energy barrier, which is smoothed by small α values. Such a role of α in controlling the roughness of the modified potential^{49,50} and the dynamics of the simulated system has been investigated in previous studies.^{35,51} In the case of POPC, when α becomes too small, the energy barrier of the double bond is significantly lowered, and the artificial *cis*–*trans* transition may occur. Such an artifact (Figure S4, Supporting Information) is observed in three aMD simulations, E2a1, E2a2, and E3a1, where $d\alpha$ was set to 1, 3, and 10 kcal/mol/lipid, respectively. These aMD simulations were terminated at $t = 10$ ns and not included in any further analysis. Apart from simulations designed to study the double bond *cis*–*trans* transition, small α values should be used with caution to avoid similar artifacts in aMD simulations of unsaturated fatty acids.

Another interesting observation from Figure 2 is that once the acceleration has been raised to a certain level, either through increasing E or decreasing α , the effect of further acceleration can be very limited. For instance, while the POPC E3a2 simulation has the largest E and the smallest α , its lateral diffusion coefficient is very similar to E2a3, which has the second highest acceleration level among aMD simulations of POPC. This result may reflect the limit of the aMD method used in these simulations, where only the dihedral potential is boosted (the aMDd mode). The speedup observed in these simulations may be attributed to a more flexible lipid structure, as revealed by the faster *trans*–*gauche* isomerization of lipid tails (Figure S2, Supporting Information) and increased rotation and barrier-crossing events of the headgroup (Figure S5, Supporting Information). Since nonbonded interactions are known to play an important role in bilayer dynamics, the lipid lateral diffusion can be expected to be further enhanced using aMDT (boosting the total potential) or aMDdual (dual boost) simulations, where the boost potential is applied to all degrees of freedom in the system.^{35,36} As a test of this hypothesis, we performed an aMDdual simulation on the POPC bilayer (Figure S6, Supporting Information), where the same boost potential used in the E3a2 simulation was chosen for the dihedrals, with a separate boost potential ($dE = d\alpha = 0.15$ kcal/mol/atom, see Figure S6) applied on the remaining

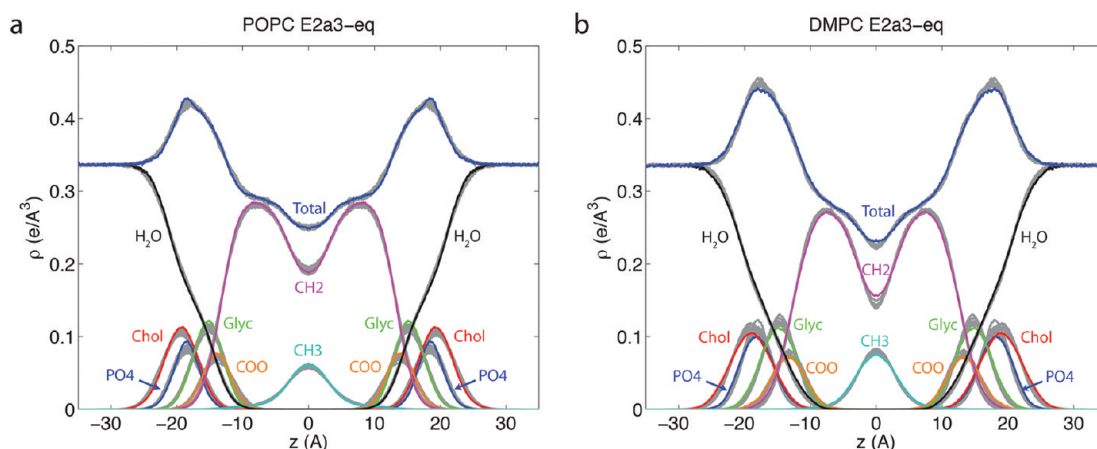


Figure 4. Electron density profiles of POPC (a) and DMPC (b) in selected aMD-eq simulations. Calculations were performed using the last 5 ns of the 10-ns aMD-eq simulations (colored lines). For comparison, the same calculations were repeated using a 5-ns block for the last 20 ns of the cMD1 and cMD2 simulations (eight gray lines).

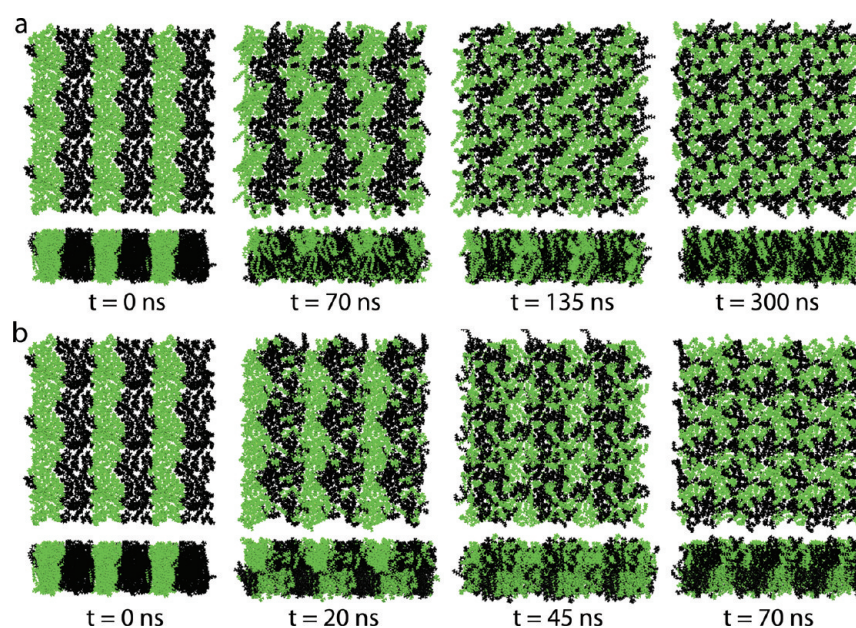


Figure 5. Representative snapshots of the mixed POPC:DMPC bilayer in cMD (a) and aMD (b) simulations. The POPC and DMPC molecules are colored in black and green, respectively. For clarity, only one monolayer is displayed in the top view figures, and nine periodic images, including the original unit cell in the middle, are shown. Both monolayers are included in the side view figures, and three periodic images are shown.

degrees of freedom. As shown in Figure S6, the lateral diffusion is enhanced by an additional 45% relative to the E3a2 simulation described earlier. Further optimization of the aMD parameters may lead to an even greater speedup and is currently under investigation in our lab.

Equilibration of aMD Systems. As shown in Figure S2 (Supporting Information), the *trans/gauche* ratios obtained from aMD simulations are slightly different from the corresponding cMD results. Indeed, the equilibrium *trans/gauche* ratio ranges from 1.90 to 2.04 in the POPC aMD simulations and from 1.98 to 2.31 in the DMPC aMD simulations, while the corresponding cMD values, averaged over cMD1 and cMD2, are 2.08 for POPC and 2.44 for DMPC. Such differences are expected for these unweighted aMD simulations, since the ensemble average of an observable needs to be recovered from

the reweighting procedure given in eq 3. Unfortunately, straightforward application of eq 3 remains a challenge for systems with tens of thousands of atoms, due to the statistical noise associated with the exponential form of the reweighting equation, which tends to manifest any small fluctuations in the boost potential.⁵²

In this work, we explored the use of a short cMD simulation to bring the aMD system back to the original, unbiased ensemble. As described in the Methods section, we performed a 10-ns cMD simulation seeded from the structure with the highest boost potential in the last 5 ns of an aMD trajectory. The high boost potential determines that the corresponding structure has a relatively large weight in the reweighting process. From the aMD-eq simulations, we calculated various structural properties of the bilayer, including the order parameter S_{CD} and the electron

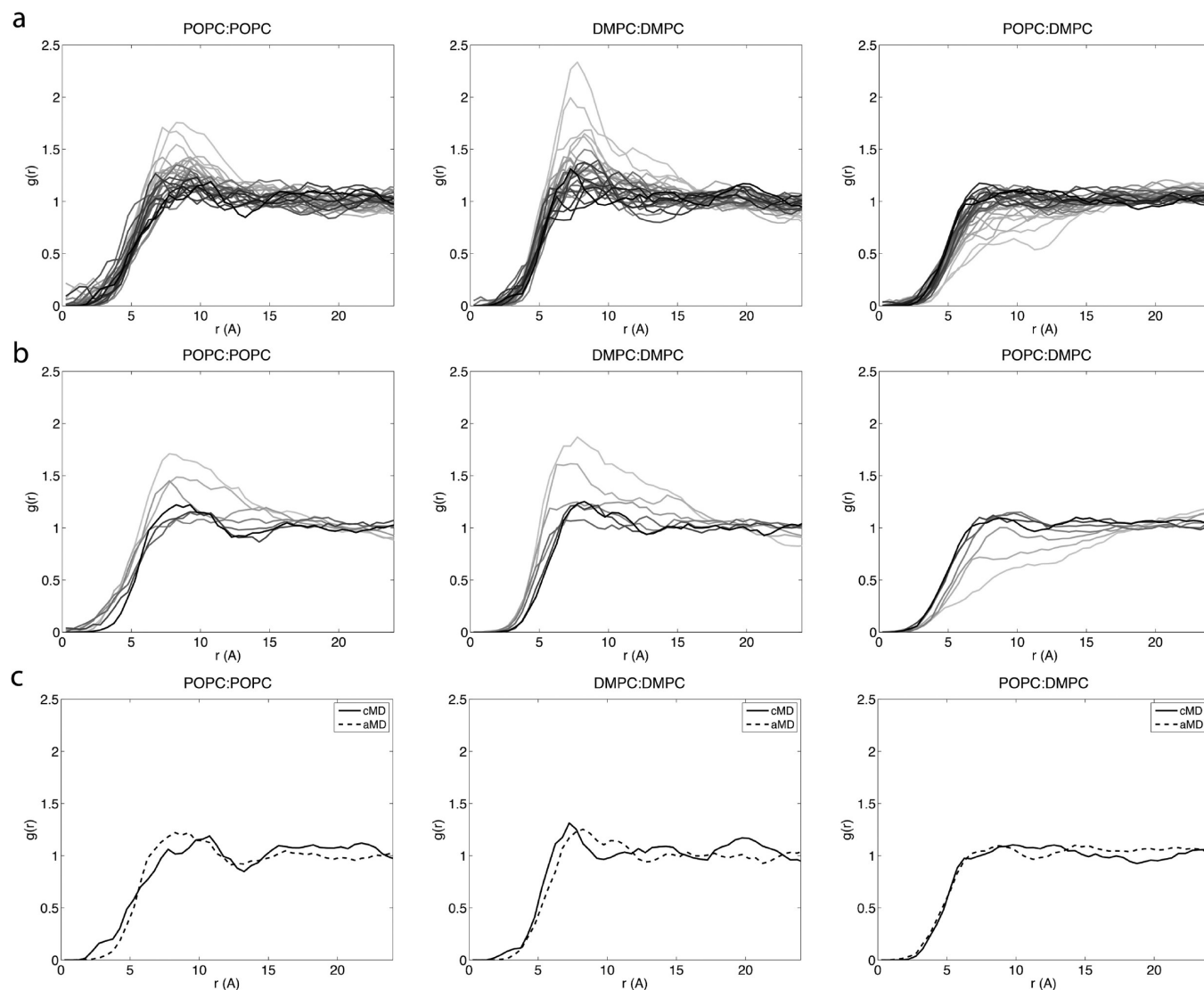


Figure 6. Radial pair distribution functions of the mixed POPC:DMPC bilayer. (a, b) Results of the cMD (a) and aMD (b) simulations. The $g(r)$ calculations were performed using a 10-ns block for three lipid pairs—POPC:POPC, DMPC:DMPC, and POPC:DMPC. Results are colored in light gray for the beginning of a simulation and dark gray for the end of the simulation. (c) Comparison of $g(r)$ obtained from the last 10 ns of the cMD (solid line) and aMD (dashed line) simulations.

density profile (EDP). By comparing these properties with results from the 70-ns cMD simulations, we evaluated the ability of the aMD-eq simulations to bring the bilayer systems to the original potential energy surface.

The deuterium order parameter S_{CD} , which is a measure of the disorder in lipid tails, was calculated from our simulations according to

$$S_{CD} = \frac{1}{2} \langle 3 \cos^2 \theta - 1 \rangle \quad (5)$$

where θ is the angle between the CH-bond vector and the bilayer normal. The electron density profile EDP was obtained according to Feller et al.,⁵³ namely, the time-averaged number of electrons was counted for every 0.1-Å slab along the membrane normal. As shown in Figures 3 and 4, overall, the S_{CD} and EDP calculated from the aMD-eq simulations are very similar with the cMD results. A slightly larger deviation is observed for the S_{CD} of the POPC sn-2 chain in the E3a2-eq simulation (Figure 3).

In light of this deviation, we extended the E3a2-eq run to 20 ns and repeated the S_{CD} calculations. As shown in Figure S7 (Supporting Information), the S_{CD} results of the extended simulation agree well with the cMD data. Additionally, we also found good agreement between aMD-eq and cMD results for the distribution profiles of all headgroup dihedrals, as well as the equilibrium *trans*–*gauche* ratios of lipid tails (data not shown).

On the basis of the above analysis, we conclude that the 10-ns aMD-eq simulations allowed most bilayer systems to relax back to the original, unbiased ensemble. However, since longer equilibration was required for the E3a2 simulation of POPC, these results also suggest that aggressive acceleration levels should be used with caution, since they might render undesirable structural artifacts or require longer equilibrations following aMD. Overall, the relatively short equilibrations described above suggest that our bilayers were not driven too far away from the original ensemble by the aMD boost potential. This result may be attributed to the relatively modest acceleration applied in the

current study, as well as the homogeneous nature of a lipid bilayer, which tends to have a smoother potential surface compared with protein systems. A number of studies have investigated post-aMD analysis methods for protein systems^{28,29,54} and the reweighting issue described earlier.^{55,56}

Mixing of POPC and DMPC. Encouraged by the enhanced lipid lateral diffusion in aMD simulations, we set out to examine the effect of aMD on lipid mixing. On the basis of the lateral diffusion calculation described above, we set the acceleration level to $dE = 37.6$ kcal/mol/lipid and $d\alpha = 18.8$ kcal/mol/lipid, equivalent to the E3a2 simulations of pure POPC or DMPC, which produced the best performance in enhancing lipid diffusion without affecting the conformation of the POPC sn-2 chain.

As shown in Figure 5, the mixing of POPC and DMPC is significantly expedited by the aMD simulation. To quantify the comparison, we calculated the 2D radial distribution functions, $g(r)$, for the lipid pairs DMPC:DMPC, POPC:POPC, and POPC:DMPC. The calculation was performed for each monolayer separately, using the projection of lipid COMs on the membrane plane. The evolution of $g(r)$ in both cMD and aMD simulations is highlighted by the color change from light gray to dark gray in Figure 6. As clearly shown in this figure, during both simulations, the $g(r)$ peak at $r = 8-10$ Å in the DMPC:DMPC and POPC:POPC pair distribution functions is gradually smoothed, indicating that the lipid molecules are no longer characterized by the clustering of like neighbors. The $g(r)$ of the POPC:DMPC pair has an opposite trend, which corresponds to an increasing degree of mixing. Note that the final $g(r)$ functions are very similar in the two simulations (Figure 6c), which indicates that the aMD result is in good agreement with the cMD simulation.

Analysis of the $g(r)$ data shows that the lipids are well mixed at $t = 50$ ns in the aMD simulation, which is reflected in a small (0.21) root-mean-square-deviation (RMSD) of the POPC:DMPC $g(r)$ compared to the final $g(r)$. In contrast, the RMSD of the POPC:DMPC $g(r)$ is 0.77 at $t = 70$ ns in the cMD simulation and only dropped to 0.29 at $t = 120$ ns. Similarly, with reference to the final DMPC:DMPC $g(r)$, the RMSD of the aMD run reached 0.33 at $t = 40$ ns, while the RMSD of the cMD run is 0.43 at $t = 130$ ns. On the basis of these results, we estimate that aMD affords an approximate 2–3-fold speedup in lipid mixing compared to the cMD simulation.

Interestingly, both the cMD and aMD results suggest that the POPC:DMPC bilayer may deviate from an ideal mixture, which is in agreement with phase diagrams derived from calorimetric data.^{57,58} As shown in Figure 5, small clusters of like lipids can be identified at the end of both aMD and cMD simulations. Meanwhile, the $g(r)$ plots revealed a small peak at $r = 8-10$ Å in the final DMPC:DMPC and POPC:POPC radial distribution functions, which is absent in the POPC:DMPC result (Figure 6c). These data reflect a tendency for lipids of the same species to aggregate in the mixed POPC:DMPC bilayer, which may be explained by nonideal mixing of the bilayer. However, even for an ideal mixture, the system is likely to experience fluctuations and occasionally deviate from ideality. Therefore, cMD or aMD simulations of longer duration are needed to fully understand the mixing behaviors of the POPC:DMPC system at the atomistic level. Given the lateral diffusion results discussed earlier, larger bilayers of mixed lipid species may be used to reduce the finite size effect.⁴⁷ The computational resources conserved through the use of aMD may be even greater for

these larger bilayers, since the cost of a simulation is proportional to the size of the system.

CONCLUSIONS

Using 1.5- μ s simulations, we studied the effect of accelerated MD on *trans*–*gauche* isomerization, lateral diffusion, and lipid mixing of three bilayer systems—POPC, DMPC, and mixed POPC:DMPC. Overall, aMD produced a significant speedup in lipid equilibration and diffusion: For the pure POPC and DMPC bilayers, aMD produced up to 8 times faster *trans*–*gauche* isomerization and up to a 3-fold speedup in lipid lateral diffusion. From a comparative analysis of seven aMD simulations for each of the POPC and DMPC bilayers, we examined the effect of aMD parameters on the structural and dynamic properties of the two lipid species. POPC was found to be more sensitive to the acceleration factor α , which controls the shape of the modified potential energy surface. We demonstrate that small α values can produce significant speedup in lipid lateral diffusion. However, on further decreasing α below a certain threshold, the artificial *cis*–*trans* transition may occur in the POPC oleoyl chain, due to the smoothed energy barrier associated with the double bond. For this reason, α should be chosen with caution to avoid such an artifact in aMD simulations of unsaturated lipids. Meanwhile, selective aMD,⁵⁶ where the boost potential is only applied to a certain part of the system, may prove a useful alternative for these systems.

On the basis of the results of pure POPC and DMPC bilayers, we tested the ability of aMD to enhance lipid mixing in the POPC:DMPC bilayer. A 70-ns aMD simulation and a 300-ns cMD run revealed similar mixing behaviors, with aMD producing a 2–3-fold increase in the mixing speed. Since interactions with lipids are crucial to the stability and proper function of a large number of membrane proteins, the aMD method may be particularly useful in speeding up the equilibration of various lipid species surrounding a membrane protein. The results presented in this work provide the benchmarks for these future aMD studies of bilayers with single- or multiple-lipid components. Additionally, the parameters examined here can be used as a starting point for further optimization in aMD simulations of membrane systems.

ASSOCIATED CONTENT

S Supporting Information. Probability distribution of lipid tail dihedrals (Figure S1), the *trans/gauche* ratio (Figure S2), the mean square displacement of lipid molecules (Figure S3), the conformations of the sn-2 oleoyl chain in POPC (Figure S4), barrier-crossing events (Figure S5), and the mean square displacement in dual-boost aMD simulation (Figure S6). This information is available free of charge via the Internet at <http://pubs.acs.org/>

AUTHOR INFORMATION

Corresponding Author

*E-mail: yiwang@ucsd.edu.

ACKNOWLEDGMENT

We thank Dr. Richard Pastor for suggesting the lipid mixing simulations. This work has been supported in part by the National Science Foundation, the National Institutes of Health,

Howard Hughes Medical Institute, Center for Theoretical Biological Physics, the National Biomedical Computation Resource, and the NSF supercomputer centers. The authors appreciate the computational resources provided by the Texas Advanced Computing Center (TG-MCA93S013 and TG-MCB090187).

REFERENCES

- (1) Feller, S. E. *Curr. Opin. Colloid Interface Sci.* **2000**, *5*, 217–223.
- (2) Saiz, L.; Bandyopadhyay, S.; Klein, M. L. *BSR* **2002**, *22*, 151–173.
- (3) Pastor, R. W.; Venable, R. M.; Feller, S. E. *Acc. Chem. Res.* **2002**, *35*, 438–446.
- (4) Lopez, C.; Moore, P.; Shelley, J.; Shelley, M.; Klein, M. *Comput. Phys. Commun.* **2002**, *147*, 1–6.
- (5) McWhirter, J. L.; Ayton, G.; Voth, G. A. *Biophys. J.* **2004**, *87*, 3242–3263.
- (6) Ayton, G. S.; Voth, G. A. *Biophys. J.* **2004**, *87*, 3299–3311.
- (7) Klauda, J.; Brooks, B.; Pastor, R. *J. Chem. Phys.* **2006**, *125*, 144710.
- (8) Klauda, J.; Kucerka, N.; Brooks, B.; Pastor, R.; Nagle, J. *Biophys. J.* **2006**, *90*, 2796–2807.
- (9) Izvekov, S.; Voth, G. A. *J. Chem. Theory Comput.* **2006**, *2*, 637–648.
- (10) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.
- (11) Berkowitz, M. *Biochim. Biophys. Acta* **2009**, *1788*, 86–96.
- (12) Falck, E.; Róg, T.; Karttunen, M.; Vattulainen, I. *J. Am. Chem. Soc.* **2008**, *130*, 44–45.
- (13) Lindahl, E.; Edholm, O. *Biophys. J.* **2000**, *79*, 426–433.
- (14) Saiz, L.; Klein, M. L. *Acc. Chem. Res.* **2002**, *35*, 482–489.
- (15) Pitman, M. C.; Suits, F.; Gawrisch, K.; Feller, S. E. *J. Chem. Phys.* **2005**, *122*, 244715.
- (16) Roark, M.; Feller, S. E. *J. Phys. Chem. B* **2009**, *113*, 13229–13234.
- (17) Huang, J.; Swanson, J.; Dibble, A.; Hinderliter, A.; Feigenson, G. *Biophys. J.* **1993**, *64*, 413–425.
- (18) Chiu, S.; et al. *Biophys. J.* **1999**, *77*, 2462–2469.
- (19) Smondyrev, A.; Berkowitz, M. *Biophys. J.* **1999**, *77*, 2075–2089.
- (20) Sugár, I. P.; Thompson, T. E.; Biltonen, R. L. *Biophys. J.* **1999**, *76*, 2099–2110.
- (21) Pandit, S.; Bostick, D.; Berkowitz, M. *Biophys. J.* **2004**, *86*, 1345–1356.
- (22) Zhang, Z.; Bhide, S.; Berkowitz, M. *J. Phys. Chem. B* **2007**, *111*, 12888–12897.
- (23) Coppock, P. S.; Kindt, J. T. *Langmuir* **2009**, *25*, 352–359.
- (24) Yin, F.; Kindt, J. T. *J. Phys. Chem. B* **2010**, *114*, 8076–8080.
- (25) Hamelberg, D.; Mongan, J.; McCammon, J. *J. Chem. Phys.* **2004**, *120*, 11919–11929.
- (26) Markwick, P.; Bouvignies, G.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 4724–4730.
- (27) Markwick, P.; Bouvignies, G.; Salmon, L.; McCammon, J.; Nilges, M.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968–16975.
- (28) Grant, B.; Gorfe, A.; McCammon, J. *PLoS Comput. Biol.* **2009**, *5*, e1000325.
- (29) Bucher, D.; Grant, B.; Markwick, P.; McCammon, J. *PLoS Comput. Biol.* **2011**, *7*, e1002034.
- (30) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (31) Kucerka, N.; Tristram-Nagle, S.; Nagle, J. F. *J. Membr. Biol.* **2005**, *208*, 193–202.
- (32) Petrache, H. I.; Tristram-Nagle, S.; Nagle, J. F. *Chem. Phys. Lipids* **1998**, *95*, 83–94.
- (33) Kucerka, N.; Liu, Y.; Chu, N.; Petrache, H. I.; Tristram-Nagle, S.; Nagle, J. F. *Biophys. J.* **2005**, *88*, 2626–2637.
- (34) Wang, Y.; Harrison, C. B.; Schulten, K.; McCammon, J. A. *Comput. Sci. Discovery* **2011**, *4*, 015002.
- (35) de Oliveira, C.; Hamelberg, D.; McCammon, J. *J. Phys. Chem. B* **2006**, *110*, 22695–22701.
- (36) Hamelberg, D.; de Oliveira, C.; McCammon, J. *J. Chem. Phys.* **2007**, *127*, 155102.
- (37) Hamelberg, D.; Shen, T.; McCammon, J. *J. Am. Chem. Soc.* **2005**, *127*, 1969–1974.
- (38) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D., Jr.; Pastor, R. W. *J. Phys. Chem. B* **2010**, *114*, 7830–7843.
- (39) Klauda, J.; Brooks, B.; MacKerell, A. D.; Venable, R.; Pastor, R. *J. Phys. Chem. B* **2005**, *109*, 5300–5311.
- (40) Klauda, J.; Pastor, R.; Brooks, B. *J. Phys. Chem. B* **2005**, *109*, 15684–15686.
- (41) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (42) Andersen, H. C. *J. Chem. Phys.* **1983**, *52*, 24–34.
- (43) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1993**, *13*, 952–962.
- (44) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (45) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
- (46) Moore, P. B.; Lopez, C. F.; Klein, M. L. *Biophys. J.* **2001**, *81*, 2484–2494.
- (47) Yeh, I.-C.; Hummer, G. *J. Phys. Chem. B* **2004**, *108*, 15873–15879.
- (48) de Oliveira, C.; Hamelberg, D.; McCammon, J. *J. Chem. Phys.* **2007**, *127*, 175105.
- (49) Hamelberg, D.; Shen, T.; McCammon, J. *J. Chem. Phys.* **2006**, *125*, 094905.
- (50) Johnson, Q.; Doshi, U.; Shen, T.; Hamelberg, D. *J. Chem. Theory Comput.* **2010**, *6*, 2591–2597.
- (51) Doshi, U.; Hamelberg, D. *J. Chem. Theory Comput.* **2011**, *7*, 575–581.
- (52) Shen, T.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.
- (53) Feller, S. E.; Venable, R. M.; Pastor, R. W. *Langmuir* **1997**, *13*, 6555–6561.
- (54) Markwick, P.; Cervantes, C.; Abel, B.; Komives, E.; Blackledge, M.; McCammon, J. *J. Am. Chem. Soc.* **2010**, *132*, 1220–1221.
- (55) Fajer, M.; Swift, R.; McCammon, J. *J. Comput. Chem.* **2009**, *30*, 1719–1725.
- (56) Wereszczynski, J.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 3285–3292.
- (57) Davis, P.; Keough, K. *Chem. Phys. Lipids* **1984**, *35*, 299–308.
- (58) Curatolo, W.; Sears, B.; Neuringer, L. *J. Biochim. Biophys. Acta Biomembr.* **1985**, *817*, 261–270.

Molecular Monte Carlo Simulations Using Graphics Processing Units: To Waste Recycle or Not?

Jihan Kim,^{*,†} Jocelyn M. Rodgers,[‡] Manuel Athènes,[§] and Berend Smit^{*,||}

[†]Lawrence Berkeley National Laboratory, [‡]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

[§]Service de Recherches de Métallurgie Physique - CEA/Saclay, 91191 Gif-sur-Yvette, France

^{||}Department of Chemical and Biomolecular Engineering and Department of Chemistry, University of California, Berkeley, California 94720, United States

ABSTRACT: In the waste recycling Monte Carlo (WRMC) algorithm,¹ multiple trial states may be simultaneously generated and utilized during Monte Carlo moves to improve the statistical accuracy of the simulations, suggesting that such an algorithm may be well posed for implementation in parallel on graphics processing units (GPUs). In this paper, we implement two waste recycling Monte Carlo algorithms in CUDA (Compute Unified Device Architecture) using uniformly distributed random trial states and trial states based on displacement random-walk steps, and we test the methods on a methane–zeolite MFI framework system to evaluate their utility. We discuss the specific implementation details of the waste recycling GPU algorithm and compare the methods to other parallel algorithms optimized for the framework system. We analyze the relationship between the statistical accuracy of our simulations and the CUDA block size to determine the efficient allocation of the GPU hardware resources. We make comparisons between the GPU and the serial CPU Monte Carlo implementations to assess speedup over conventional microprocessors. Finally, we apply our optimized GPU algorithms to the important problem of determining free energy landscapes, in this case for molecular motion through the zeolite LTA.

1. INTRODUCTION

With the introduction of multicore chips, a new paradigm of scientific computing has emerged in which scientific application codes took advantage of on-chip parallelism provided by the hardware. As computing capabilities move toward an era of exascale computing, various hardware, such as many-core CPUs, GPUs, and CPU–GPU fusion architectures are emerging to provide the next important shift in the area of high performance computing. Originally intended to handle computation for graphics, GPU scientific computing has introduced new parallelization techniques that are being utilized in solving scientific problems. Compared to CPUs, GPUs have more transistors devoted to data processing as opposed to cache memory and loop control, and accordingly programs that can be efficiently mapped onto this multithreaded hardware can see significant performance improvement. To achieve efficient computations, the GPU and CPU can work together in a heterogeneous coprocessing computing model where the sequential part of the code can be executed inside the CPU while the computationally intensive massively parallel part of the code can be accelerated inside the GPU. Traditionally, GPUs have been used mostly for graphics intensive applications but the release of NVIDIA's CUDA has allowed programmers to use C-like syntax language for code development, extending its utility for scientific computing.² Thus far in computational chemistry, there has been substantial GPU code development in both molecular dynamics (MD)³ and Monte Carlo (MC) simulations.^{4–6} MC simulations are very similar to MD, and as such, many of the techniques developed for molecular dynamics simulations such as neighbor lists and cell lists⁷ can also be used in a Monte Carlo simulations. However, for some systems for which molecular dynamics

simulations can be extremely time-consuming or simply not feasible, special Monte Carlo algorithms can make these simulations orders of magnitude more efficient.⁷ Of particular interest are simulations of open systems, which rely on the addition or removal of particles. Such systems can be conveniently simulated in the grand-canonical ensemble using a Monte Carlo simulation. In this paper, we shall focus on techniques to accelerate the MC simulations of molecular systems on the GPU that expand beyond the acceleration present in typical GPU-based MD codes.

In essence, this paper seeks to address the broad question of how we may best leverage GPU resources in conducting *molecular Monte Carlo* simulations. Here, we consider three alternative avenues to accelerating convergence of a simulated thermodynamic property—the average energy per molecule. We then apply these acceleration strategies to the estimation of free energies, an important and difficult goal of molecular simulation. The three strategies are orthogonal and involve (1) an embarrassingly parallel implementation of many side-by-side Monte Carlo simulations, (2) the parallelization of the pairwise summation of Lennard-Jones (LJ) interactions, and (3) the use of multi-proposal Monte Carlo coupled with waste recycling on the rejected states. The first avenue, embarrassingly parallel simulations, is certainly available to both molecular dynamics and Monte Carlo simulations; however, the small number of mobile molecules in these Monte Carlo simulations make this parallelization strategy tractable. The motivation for this approach lies in using each GPU thread to gather as much independent statistics as possible. The second avenue, parallelization of the

Received: July 8, 2011

Published: September 22, 2011

pairwise summation of LJ energies, is more pertinent to MC simulations, while in principle being amenable to MD simulations. While the LJ pair potential is employed in molecular dynamics simulations, there are two distinctions in its application in these MC simulations. For MD, all particles are moved at each time step, requiring a reevaluation of all pair forces at each time step, while in MC, one particle is moved for each MC step, resulting in fewer pairs contributing to changes in interactions at each step. Furthermore, as MC is only concerned with changes in the total energy of the system, reduction on the GPU of the sum over pair interactions results in a single systemwide energy, rather than distinct forces applied to each individual particle, resulting in a simpler reduction algorithm. These factors combine to yield a simplified parallelization scheme for LJ interactions. The goal of this approach is to conduct each single, traditional Monte Carlo step as quickly as possible. The third avenue, multiproposal MC combined with waste recycling, is also available solely to MC simulations and is an option for the use of computing power which is wholly novel. In essence, the threads of the GPU are employed to propose many possible new states at each MC step, and waste recycling is employed to harvest as much information from both the chosen state and the rejected states. Waste recycling has been employed in tandem with molecular dynamics previously through the use of multiple time slice estimators.⁸ The route to parallelizing such an approach would rather focus on the energy and the force evaluations of the proposals that are generated successively by the molecular dynamics, while multiproposal MC algorithms are inherently parallelizable.

At this point, we would like to emphasize that the above techniques are generally applicable. However, as in any Monte Carlo simulation, the efficiency depends very much on the details of the system. Therefore, we illustrate the application of these techniques to a system of practical importance, methane adsorbed in the zeolites MFI (silicalite) and LTA. Zeolites are nanoporous materials important in catalysis and separations for (petro)chemical processes, and molecular simulation has proven invaluable in evaluating the thermodynamics and kinetics associated with molecular absorption and motion through zeolites.⁹ For these systems, the pores can slow down the dynamics of the adsorbed molecules by orders of magnitude, making the use of Monte Carlo techniques crucial. In addition, adsorption isotherms are conveniently calculated in the grand-canonical ensemble, which requires the use of Monte Carlo simulations.

The paper is organized as follows. In section 2, we describe the molecular modeling and Monte Carlo simulation of methane adsorbed in zeolites. In section 3, we outline the mapping of the methane–zeolite system onto a GPU. In section 4, we describe in detail our GPU Monte Carlo algorithms as well as provide optimization techniques utilized in each of the different parallelization methods. In section 5, we present results of our simulations for the methane–MFI system for the different parallelization approaches, and we discuss the merits of each approach. We also apply our GPU acceleration schemes to determining the free energy profile of methane adsorbed in the LTA zeolite. Finally, in section 6, we summarize the important issues mentioned in the paper and briefly look ahead to future work.

2. ZEOLITE SIMULATIONS

2.1. Molecular Models for Zeolites. Zeolites are nanoporous crystalline materials. At present, there are over 200 different crystal structures, each with a different pore structure. The size of

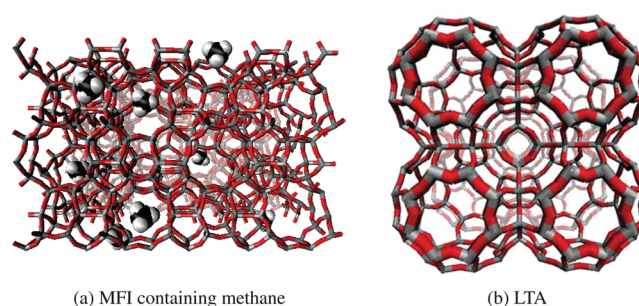


Figure 1. Snapshots of two zeolites with oxygen in red and silicon in gray. The adsorbate methane is shown only in MFI with carbon in black and hydrogen in white. Each zeolite is viewed from the (100) axis and contains a total of eight unit cells. The front four are opaque, and the back four are transparent. The LTA simulation box is composed of eight supercells connected to each other in the x , y , and z directions via windows of smaller radius and higher free energy. The high symmetry of this structure allows us to visualize these passageways relatively easily. In the snapshot of LTA, the four opaque quadrants of the zeolite correspond to the front four supercells of the zeolite, viewed through the frontmost windows shown with thicker bonds. MFI is composed of zigzag and straight channels. The 10 methane adsorbates are most visible along the (100) axis in this instantaneous configuration at locations within the zigzag channels directed along the x axis. However, the channels themselves are not visible in completion due to their kinked nature.

these pores ranges from a few angstroms to 2 nm. In this work, we study MFI (Figure 1a) and LTA (Figure 1b), which are both frequently studied materials.^{10–12}

These two zeolites present markedly different pore topologies. MFI is composed of intersecting straight and zigzag channels. The snapshot in Figure 1a shows methane molecules adsorbed within zigzag channels, and these channels are not easily visible. In contrast, LTA is composed of an intersecting structure of large cages with narrower windows connecting them in all three directions, resulting in a simple cubic lattice of cages with free energy barriers to diffusion across each window. Figure 1b clearly displays the four frontmost cages with four posterior cages in lighter coloring.

Monte Carlo simulations have been crucial in studying these zeolite systems. In assessing molecular adsorption isotherms, Monte Carlo simulations are necessary in order to allow for particle number varying with applied chemical potential via grand canonical simulations. Furthermore, Monte Carlo simulations provide a straightforward route for calculating the free energy barriers to diffusion.^{11,13} In this work, we focus on accelerating canonical MC simulations for methane in the zeolites MFI and LTA, but the lessons gleaned from this study can be broadly applicable to grand canonical simulations of this system and MC simulations of other molecular systems.

The focus of this work is on canonical (NVT) Monte Carlo simulations where the number of particles (N), volume (V), and temperature (T) remain the same throughout the simulation. Our predominant system of interest consists of methane molecules within a zeolite framework, which we assume to be rigid. The force field is parametrized using the conventional assumptions: the zeolite framework is rigid, and the interactions are dominated by the oxygen atoms in the zeolite framework. The number of methane molecules is varied for different sets of simulations while the number of framework oxygen atoms remains fixed, leading to different loadings of the framework and different effective densities of the methane molecules. In all simulations, the temperature is fixed to 300 K.

For each zeolite simulated, the simulation box is comprised of eight ($2 \times 2 \times 2$) unit cells. The three-dimensional MFI crystal unit cell has dimensions of $20.022 \text{ \AA} \times 19.899 \text{ \AA} \times 13.383 \text{ \AA}$ and contains 288 framework atoms. The MFI simulation box contains a total of 2304 (1536 oxygen and 768 silicon) atoms. The LTA crystal unit cell is cubic with a side length of 12.278 \AA . There are 72 framework atoms inside this unit cell, resulting in a total of 576 (384 oxygen and 192 silicon) atoms in the simulation box.

Interaction between methane molecules (i.e., methane–methane) and between methane molecules and the zeolite (i.e., methane–oxygen) is modeled on a pairwise basis by the Lennard-Jones potential:

$$U(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (1)$$

where r is the distance between two particles, ϵ indicates the depth of the potential well (148.0 K for methane–methane and 115.0 K for methane–oxygen),^{11,14} and σ represents the effective core size of the particles with the potential well located at $2^{1/6}\sigma$. The σ for methane–methane interactions is 3.73 \AA and that for methane–oxygen interactions is 3.47 \AA .^{11,14} In all cases, the Lennard-Jones interaction is shifted to zero for $r > R_c$ by the subtraction of $U(R_c)$ from $U(r)$ for all $r \leq R_c = 12 \text{ \AA}$.

Periodic boundary conditions are imposed, and each dimension of the simulation cell is chosen to be greater than $2R_c$. This allows the determination of various properties of the molecular system through simulation of only a small subset of the entire system. As a consequence of this and through the representation of the zeolitic framework via an energy grid as discussed in section 3.2, the Monte Carlo simulations conducted in this work consist of only a small number of molecules. The largest system contains 128 methane molecules.

2.2. Monte Carlo Simulations. Here, we give a brief summary of the Monte Carlo techniques employed, deferring discussion of most algorithmic details until section 4. For both the embarrassing parallel Monte Carlo simulations and the MC simulations using parallelized calculation of Lennard-Jones interactions, we simply conduct standard Metropolis Monte Carlo.⁷ In this approach, an adsorbate molecule is chosen at random and moved by a random displacement chosen from $[-d_{\max}, d_{\max}]$ in each direction, generating a new state n by displacement from the old state o . The change in energy ΔE is calculated, and the move is accepted or rejected by the Metropolis acceptance criterion:

$$\text{acc}_M(o \rightarrow n) = \min[1, \exp(-\beta\Delta E)] \quad (2)$$

We also implement two waste recycling Monte Carlo algorithms. The idea of using information of rejected moves in Monte Carlo simulations was first explored by Frenkel in his waste-recycling Monte Carlo scheme.¹ The principle idea of waste recycling is that all rejected states in a Monte Carlo simulation do carry some information, and when included with their proper Boltzmann weights, they can improve the statistics. Delmas and Jourdain¹⁵ have in fact proven that when MC is conducted with the symmetric Boltzmann or Barker acceptance criterion,

$$\text{acc}_B(o \rightarrow n) = \frac{\exp(-\beta E_n)}{\exp(-\beta E_o) + \exp(-\beta E_n)} \quad (3)$$

waste recycling is guaranteed to lead to faster convergence of statistical properties.

The waste-recycling method is inherently parallelizable in the sense that you may generate multiple trial states for a single Monte Carlo step, as done by Frenkel,¹ and accordingly is suited to simulate using a GPU. It is therefore interesting to investigate whether the combination of GPU with waste recycling is an attractive route for Monte Carlo simulations. The multiproposal algorithm suggested by Frenkel is most readily applied when new particle positions are chosen randomly throughout the entire simulation box, as discussed in section 4.3. However, with care, an analogue based on particle displacements may be constructed, as in section 4.4. For readers familiar with configurational bias Monte Carlo, we note that this approach for generating multiple trial states is distinct, as discussed more in Appendix .

2.3. Simulated Properties—Energy and Free Energy. For the bulk of our studies of GPU algorithms, we shall study the simple property of average energy of methane adsorbed in the zeolite MFI, as energies must be calculated at each step in the Monte Carlo simulation. Average energy is not typically a difficult quantity to converge; however, it yields a well-defined benchmark for gauging the speedup of our various GPU algorithms and to study the optimization of these algorithms on the GPU architecture.

Once we have optimized the various algorithms, we then apply the lessons gleaned to a highly relevant property, the free energy profile of methane diffusing through the zeolite LTA. In zeolites, these free-energy barriers are relevant for characterizing the diffusive behavior of adsorbates within the zeolite framework.¹⁶ The zeolite LTA poses a straightforward pore topology of a simple cubic arrangement of pores separated by windows. Thus, in any axial direction, adsorbed methane molecules encounter free energetic barriers at the windows connecting the cages. This barrier is substantial yet still thermally accessible. As such, the calculation of the free-energy profile along the reaction coordinate z is simply calculating the histogram of probabilities along z , $P(z)$:

$$F(z) = -k_B T \ln P(z) \quad (4)$$

Certainly, this examination is relevant because $F(z)$ is a meaningful and computationally intensive quantity to calculate. But this final comparison is also important to conduct because waste-recycling Monte Carlo is most fruitfully employed on CPUs for free energy calculations,^{1,8,17,18} so we must also allow for this possibility on the GPU as well.

3. COMPUTATIONAL SETUP

In this section, we first outline the various strategies we employ for mapping these zeolite simulations onto the GPU. These strategies as well as the implementation, optimization, and testing of GPU algorithms for these strategies are the main contributions of this paper, as detailed in subsequent sections. We then describe in section 3.2 one constant in all of our studied GPU implementations—the construction of a framework energy grid to represent the interactions between methane molecules and the zeolite framework atoms. This grid in general yields a substantial speedup in the simulation of zeolites, which are rigid to a good approximation; thus careful implementation on the GPU is important.

All of the simulations for this work were conducted using the Dirac cluster at the National Energy Research Scientific Computing Center (NERSC). The CURAND library was employed for generating the necessary random numbers on the GPU.¹⁹

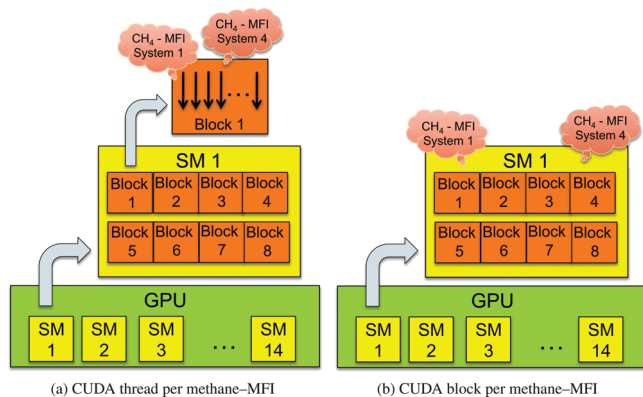


Figure 2. Mapping of methane–MFI systems to CUDA hardware resources. In Tesla C2050 GPU cards, there are 14 streaming multiprocessors (SM) shown as green blocks and at maximum eight resident blocks (Bl) per SM (shown as red boxes). Each thread within a block is indicated by a black arrow. (a) The CUDA thread per methane–MFI implementation mapping is used in the embarrassingly parallel algorithm (PMC). (b) The CUDA block per methane–MFI implementation mapping is used in the parallel Lennard-Jones algorithm (PLJ) and the two multiproposal waste recycling algorithms (MUP and MDP). As an example, we show the “location” of methane–MFI system 1 and system 4 on hardware resources for each mapping.

Further hardware and compilation details may be found in Appendix .

3.1. Strategies for Mapping Zeolite Simulations onto GPU.

Given that only a small subset of the entire structure needs to be simulated, thousands of methane–MFI systems can simultaneously fit inside the GPU DRAM and multiple, independent Monte Carlo simulations can be processed in parallel to improve the statistical accuracy. We pursue two different ways to map the individual methane–MFI system to the GPU hardware.

First, we choose a strategy where each CUDA thread executes its own independent MC simulation (Figure 2a). Denoting the CUDA block size as n_{threads} , a single Monte Carlo simulation contains $N_{\text{blocks}} \times n_{\text{threads}}$ independent methane–MFI systems. Effectively, this algorithm can process n_{threads} times more methane–MFI systems in one simulation compared to the next parallelization scheme conducted on a per block basis. Despite this benefit, this implementation includes overall longer computational wall time and more DRAM usage that reduces the maximum number of particles that can be simulated, and it also cannot utilize fast GPU memory due to the large amount of resources used by a single CUDA thread.

Therefore, we also choose an alternate parallelization strategy where each CUDA thread block executes its own independent MC simulation with threads from the same block contributing to either waste recycling or parallel Lennard-Jones calculations (Figure 2b). N_{blocks} is used to represent the total number of CUDA thread blocks launched and, in this context, refers to the total number of independent methane–MFI systems processed in a single simulation. Communications between different CUDA thread blocks is unnecessary until the end of the simulation when energy values are collected from each of the blocks to obtain an ensemble average.

3.2. Construction of the Framework Energy Grid. In order to save computation time, explicit Lennard-Jones pair potential calculation between the methane and the framework molecules is avoided. Instead, we construct a three-dimensional rectangular

grid superimposed on top of the entire simulation box, where each of the points represents the Lennard-Jones pair potential values between a single methane molecule and the entire molecular framework. All of the energy grid point values are computed just once before the start of the Monte Carlo simulation and subsequently used during the Monte Carlo simulation as a lookup table. By utilizing thousands of lightweight threads available in the GPU architecture, the energy grid point values can be computed in parallel inside the GPU using a simple domain decomposition method in which each CUDA thread is responsible for computing a single grid point value. To improve accuracy within this approximation, it is better to generate as many points as possible while working within the memory available in the GPU DRAM. With an energy grid of mesh size $512 \times 512 \times 256$ along the x , y , and z directions, results within 0.05% of the energy values from utilizing direct Lennard-Jones calculations are obtained. Therefore, we use the corresponding values of $\delta x = 0.0784 \text{ \AA}$, $\delta y = 0.0778 \text{ \AA}$, and $\delta z = 0.105 \text{ \AA}$ as our mesh size for all of the simulations.

The memory needed to store the energy grid values is too large (roughly 500 MB for 64-bit doubles) to fit into any of the fast GPU memory, and hence the energy grid array is put into the slow, global GPU DRAM. Given that the methane molecules are free to occupy spatial coordinates not directly located on the energy grid points, linear interpolation functions from eight nearest neighboring energy grid point values are used to approximate the exact Lennard-Jones pair potential value at a given particle position at each Monte Carlo step. Because these eight neighboring points cannot be stored contiguously inside the GPU memory for all of the grid points, multiple memory transactions are needed to compute the contributions of the framework molecules from a single particle position. Although these memory transactions are expensive, there still exists considerable speedup utilizing the energy grid as opposed to explicitly computing the Lennard-Jones pair potential terms from each of the 1536 framework molecules for all Monte Carlo steps. Moreover, as the number of molecules increases, the proportional wall time spent in the energy grid read access becomes smaller and becomes less of a concern. For the GPU architecture in general, frequent data movement from the GPU DRAM causes the code to become memory bound (which amounts to 144 GB/s in the Tesla C2050 cards used in our work).

Finally, the interaction between the framework molecules (i.e., oxygen–oxygen) is ignored, as oxygen is assumed to be stationary throughout the simulation.

4. GPU MONTE CARLO ALGORITHMS

In this section, we describe in detail the four GPU Monte Carlo algorithms implemented in our work. In all of the methods, one initial methane–MFI system is generated by assigning randomized positions to methane molecules inside the simulation box. This configuration is equilibrated via a serialized Markov-Chain Monte Carlo (MCMC) method using a single core CPU. In this MCMC simulation, a small (that is, small compared to the dimensions of the simulation box), random translation step size is used to propose a particle movement to a new position, and this proposal is either accepted or rejected according to the Metropolis probability.²⁰ The step size in particle displacement is chosen such that approximately 50% of the translation proposals are accepted according to the Metropolis probability. In this work, we do not focus on accelerating the equilibration phase of

the Monte Carlo, and therefore the CPU rather than the GPU is utilized to equilibrate the system. After equilibration, the particle coordinates are duplicated N_{blocks} times (in the embarrassingly parallel algorithm, $N_{\text{blocks}} \times n_{\text{threads}}$) in the CPU, and the data are transferred to the GPU DRAM. In general, we want N_{blocks} to be a number that is an integer multiple of the total number of streaming multiprocessors found in the GPU (e.g., 14 in the case of Tesla C2050) to balance the workload among the multiprocessors. Inside the GPU, these systems need to be decorrelated in order to remove any correlation that might persist between particles from different systems, which can adversely affect the MC results. The algorithm used for decorrelation is the same as that employed in the accumulation phase. This algorithm is unique to each method, and all are detailed next.

4.1. Embarrassingly Parallel Monte Carlo (PMC). In contrast to the following three methods, in the embarrassingly parallel Monte Carlo (PMC) algorithm, each CUDA thread (instead of each CUDA block) is mapped to one methane–MFI system, and all of the threads conduct their own independent monoproposal MCMC simulation. Within this implementation, since each CUDA thread needs to have its own unique data of particle positions as well as other hardware resources, the fast memory available from the GPU hardware is insufficient, and all of the data are kept inside the global DRAM. We limit global DRAM transactions via memory coalescing and utilize the following indexing scheme to store particle coordinates inside our arrays. For a given CUDA thread j , the coordinates of the particles are stored inside an array in indices j , $j + n_{\text{threads}}N_{\text{blocks}}$, $j + 2n_{\text{threads}}N_{\text{blocks}}$, etc. such that a single memory transaction executed by a *warp* (corresponding to 32 independent methane–MFI system) can load a contiguous block of 32 64-bit data related to one particle index in order to maximize memory throughput. Within this implementation, all of the CUDA threads (and therefore, each independent methane–MFI system) choose the same particle index number for displacement in a Monte Carlo step to avoid warp divergence. Once the systems are completely decorrelated from one another, choosing the same index does not cause a problem since particles that possess the same index number from different systems are unrelated to one another.

4.2. Parallel Lennard-Jones (PLJ). For most system sizes, the bottleneck routine in our Monte Carlo simulation is the kernel that computes the Lennard-Jones pair potential. Accordingly, we devise a GPU algorithm that parallelizes this calculation in the monoproposal MCMC algorithm. The change in total energy for moving particle k from $\mathbf{r}_{k,\text{old}}$ to $\mathbf{r}_{k,\text{new}}$ while holding the remaining particles fixed can be calculated in $\mathcal{O}(N_{\text{tot}})$ time via

$$\begin{aligned} \Delta E_{\text{tot}} = & \sum_{j \neq k}^{N_{\text{tot}}} U(r_{jk,\text{new}}) - \sum_{j \neq k}^{N_{\text{tot}}} U(r_{jk,\text{old}}) \\ & + E_{\text{grid}}(\mathbf{r}_{k,\text{new}}) - E_{\text{grid}}(\mathbf{r}_{k,\text{old}}) \end{aligned} \quad (5)$$

where N_{tot} is the total number of particles in the system and $U(r_{jk})$ is the Lennard-Jones pair potential between particles j and k as defined in eq 1. $E_{\text{grid}}(\mathbf{r}_k)$ represents the total summation of pair potentials between particle k and all of the framework molecules as computed by an energy grid. The computation of these energies requires only linear interpolation from given grid points, as described in the previous section. The most expensive computation is the calculation of all LJ pair energies, and for a given Lennard-Jones pair-potential calculation between two particles, the most expensive operation is the floating-point

division operator. In order to reduce the cost, only one division operator is executed per pair potential, and the intermediate term is reused to avoid the second division operation in eq 1.

In the PLJ algorithm, the CUDA threads within the same block divide up the computation work of the two summation terms found in eq 5. Energy subtotals from each of the CUDA threads are combined at each MC step using a reduction kernel to obtain the total Lennard-Jones potential value. We utilize a reduction kernel similar to one found in the CUDA SDK example.²¹

4.3. Waste-Recycling with Multiple Uniform Proposals (MUP). The previous two implementations are based on the conventional Monte Carlo algorithm and in which we use the GPU to speed up the bottleneck routines in the computation. In this section, we develop an alternative approach in which we implement the idea of using information of rejected moves in Monte Carlo simulations, using the GPU to generate multiple trial states and employing a multiproposal version of the waste recycling Monte Carlo described in section 2.2. In practice, the waste recycling algorithm can be easily mapped into conventional CPU architectures, but in the case of single-core architectures, the parallelism would be lost as each of the multiple proposed trial states in the Monte Carlo algorithm would need to be processed sequentially. On multicore CPU architectures, the different trial states can be processed in parallel, but the performance gain will not be as great compared to the execution via GPU architecture due to the larger overhead cost of generating and combining multiple CPU threads.

In our CUDA waste recycling Monte Carlo code, we first employ a variant of the waste-recycling algorithm outlined in the paper by Frenkel,¹ which we shall refer to as multiple uniform proposals (MUP). In this paper, he describes a symmetric Boltzmann-like multiproposal scheme where the original state o is included on equal footing with all of the states in the set $\{n\}$ of trial states, and the final state is chosen from the set $\{o, \{n\}\}$.

Our CUDA waste recycling algorithm based on that of Frenkel¹ is as follows:

- 1 Generate an initial state o , and set to zero the accumulator S_A for estimating the average of the observable A .
- 2 Generate a set of trial states, $\{n\}$, by randomly choosing one methane molecule and randomly generating a set of new positions for this molecule uniformly throughout the entire simulation box. The total number of trial states including the old state o is represented by N_{prop} .
- 3 Compute the Boltzmann weights $w_i = \exp(-\beta E_i)$ for all $i \in \{o, \{n\}\}$.
- 4 Update the accumulator S_A according to the following equation:

$$S_A \rightarrow S_A + \frac{\sum_i w_i A_i}{\sum_i w_i}$$

where again $i \in \{o, \{n\}\}$.

- 5 Choose a final state $f \in \{o, \{n\}\}$ with an acceptance probability

$$p_{\text{acc}}(f) = \frac{w_f}{\sum_i w_i}$$

- 6 Repeat steps 2–5 starting from the chosen state f . Continue for a total of M Monte Carlo steps.

7 Estimate the average of A as $\langle A \rangle_{\text{est}} = S_A/M$.

For this work, the accumulator term S_A defined in step 1 either refers to the total energy of the system or the occupation probability of a given volume slice. These individual probabilities for volume slices that span the simulation box are then combined in the end to yield $P(z)$ for free energy calculations.

For step 2, the total number of trial states, N_{prop} , is set to be a multiple of 32 because the NVIDIA GPU hardware schedules and executes threads in groups of 32 called a *warp*. For any other number, some threads will remain idle while waiting for other threads within the same warp to finish work, resulting in warp divergence and performance degradation. We note that step 2 is the point of main difference with our alternative multiproposal waste recycling algorithm described in the following subsection.

The total energy of proposal i , E_i , is used to determine the associated Boltzmann weight, w_i , in step 3 and is calculated from eq 5. All N_{prop} threads share the same value for the second summation term in eq 5, $\sum_{j \neq k}^{N_{\text{tot}}} U(r_{jk, \text{old}})$, and accordingly this term needs to be only calculated once for all of the multiple proposals. In our implementation, thread 0 calculates this term. This offers an advantage over a monoproposal Lennard-Jones algorithm in which the term needs to be calculated for each new displacement, and its value cannot be reused for additional displacements.

In step 4, the accumulator term S_A is updated at each Monte Carlo step by taking the summation of the energy contributions from N_{prop} CUDA threads. Two separate reduction operations need to be performed to compute $\sum_i w_i A_i$ and $\sum_i w_i$, and we utilize an algorithm similar to one found in the reduction for parallel Lennard-Jones calculations. It is noteworthy to point out that array elements with nonzero indices contain partial summation results that are later used in step 5 to expedite calculation in its routine. As one example, $w[1] = \sum_{i=0}^{N_{\text{prop}}/2} w[2i + 1]$. Because multiple accesses to the same memory address occur as many as $\log N_{\text{prop}}$ times in the reduction algorithm, all of the array elements $w_i A_i$ and w_i are fetched once from the global memory and moved into the fast, shared memory in order to reduce global memory bandwidth. Using shared memory in the reduction kernel is more important in pre-Fermi cards, which does not have the L1 cache memory. In the Fermi GPUs, the array can be kept inside the global memory and moved into the 16/48 kB L1 cache, which is the same hardware as the shared memory, avoiding performance degradation.

In step 5, the WRMC algorithm updates the particle coordinates by selecting a final state among all proposed trial states with probability proportional to its Boltzmann weight. This acceptance probability is a multiproposal extension of the symmetric Barker acceptance ratio defined in section 2.2.^{1,15} For thread 0, the final state is set to be equal to the initial state, thereby making it possible for the system to remain unchanged after a WRMC step. Using CUDA, step 5 can be conducted in $\log N_{\text{prop}}$ steps by reusing intermediate results from the step 4 reduction kernel in a following way. At the end of the reduction kernel, the array element $w[0]$ contains the sum of all of the Boltzmann weights $S = \sum_{i=0}^{N_{\text{prop}}} w_i$, whereas other array elements have partial sums that are less than S . In step 5, all of the elements in this array are divided by S for normalization purposes, such that $w[0] = 1$ and $0 < w[i] < 1$ for all $i \neq 0$. Next, a random number R uniformly distributed from 0 and 1 is generated per system using the `curand_uniform_double` function from the device API. Depending on R , the final state is

chosen according to the algorithm outlined in the CUDA code below:

Listing 1. Step 5 CUDA Code

```
1 //R - random number uniformly chosen
  from [0, 1]
2 //w - array that contains normalized Boltzmann weights
3 //step_num - index of trial state chosen for next MC
  move
4 //N - total number of independent WRMC proposals
5
6 int index = 1;
7 int N = blockDim.x;
8
9 for(int i = 2; i < N; i*=2)
10 {
11 if (R < w[index])
12 index += i;
13 else
14 {
15 R -= w[index];
16 index += 0.5*i
17 }
18 }
19
20 //last step
21 if (R >= w[step_num])
22 step_num -= 0.5*N;
```

The system updates to a new state, and the waste recycling Monte Carlo step is repeated M steps (step 6). Finally, the average quantity A for the system is obtained after M moves according to step 7.

4.4. Waste-Recycling with Multiple Displacement Proposals (MDP). The waste-recycling algorithm presented in the preceding subsection relies on nonlocal moves. Such a scheme works well if the rejected configurations have a reasonable contribution. In dense systems, however, the probability that a nonlocal move gives a significant contribution is extremely low. Therefore, we expect that local moves constructed on the basis of *displacements* from the old state, as done for our PMC, PLJ, and benchmark serial CPU algorithms, will be much more fruitful. This, however, requires a modification of the original algorithm in order to construct a waste-recycling algorithm based on multiple particle displacements from the old state. As will be demonstrated, this method is more effective at exploring important regions of phase space at high particle densities.

The only needed modification of the previous algorithm lies in constructing the set of proposals $P \equiv \{o, \{n\}\}$ in step 2, where small displacement steps are used to generate the multiple trial states in our algorithm. In generating this set P of positions based on displacements, it is *crucial* that the set is equally likely to be generated by any trial position in the set. Provided that the generation probability $p_{\text{gen}}(P|i)$ of set P is equal for any point $i \in P$, the simple Barker-like acceptance probability employed in section 4.3 may still be used.

The crux of generating such a set of positions P lies in constructing two separate random walks starting from the chosen particle k in state o . The total length of the two random walks is $N_{\text{prop}} - 1$ displacements, leading to a total of $N_{\text{prop}} - 1$ trial states. However, by choosing the position of the generating point from state o uniformly within the combined length of the two random walks, any trial position within the random walk is

equally likely to have generated the set of positions P . Therefore $p_{\text{gen}}(P|i)$ is equal for all trial states in the set, yielding the simple acceptance probability given in section 4.3 by detailed balance.

In detail, each random walk is initiated from the position of particle k in state o . The two walks are of length $N_{\text{prop}} - l - 1$ and l , with l representing a random integer from 0 to $N_{\text{prop}} - 1$. Each point in each random walk is based on a uniformly generated random displacement from the previous position in the random walk. We make sure that each of these random displacements abide by the periodic boundary condition given that they can fall outside of our simulation volume. We emphasize the exact algorithm we use for trial state generation by displacements because several seemingly reasonable alternatives for generating multiple proposals either result in incorrect sampling of states or a prohibitively expensive use of waste, as is discussed in Appendix

4.5. Optimal Estimation for Waste Recycling. The work of Delmas and Jourdain¹⁵ suggests an optimal reweighting of waste recycling averages and traditional MC averages when employing symmetric acceptance ratios such as those employed in sections 4.3 and 4.4. To our knowledge, the work presented here is the first implementation of this optimal estimator in a multiproposal framework, though this estimator has been explored in detail in a monoproposal setting.²²

For a quantity A , detailed mathematical analysis by Delmas and Jourdain demonstrated that the optimal estimation of its average, when using a Barker-like acceptance ratio, may be written as

$$\langle A \rangle_{\text{opt}} = (1 - b^*) \langle A \rangle_{\text{MC}} + b^* \langle A \rangle_{\text{WRMC}} \quad (6)$$

where optimal estimation is in the sense of minimal standard deviation. The coefficient b^* itself depends on the variance of the property as well as the correlation in a property across steps in the Markov chain as

$$b^* = \frac{\langle A^2 \rangle - \langle A \rangle^2}{\frac{1}{2} \langle (A_{m+1} - A_m)^2 \rangle} \quad (7)$$

with the A_m and A_{m+1} in the denominator indicating quantity values at successive states in the Markov chain.^{15,22} This estimator of Delmas and Jourdain is still valid when using Metropolis-like acceptance ratios such as the one described for configurational-bias MC in Appendix ; however, in such cases, the estimator is not optimal.

Since each quantity in the above equation for b^* is a pure ensemble average, these averages may be evaluated with equal validity as waste-recycling estimates and as traditional Monte Carlo averages. Thus, when waste-recycling provides a benefit, which we expect for this multiproposal scenario, a waste-recycling estimation of b^* makes sense. In such a case, the coefficient for the optimal estimator b^* itself may be estimated as

$$b^*_{\text{WRMC}} \approx \frac{\frac{1}{M} \sum_{m=1}^M \sum_{i \in \{o, \{n\}\}} p_{\text{acc}}(i) A_i^2 - \left(\frac{1}{M} \sum_{m=1}^M \sum_{i \in \{o, \{n\}\}} p_{\text{acc}}(i) A_i \right)^2}{\frac{1}{2M} \sum_{m=1}^M \sum_{i \in \{o, \{n\}\}} p_{\text{acc}}(i) (A_i - A_o)^2} \quad (8)$$

Note that the denominator in the above equation no longer references the m th and the $(m + 1)$ th states. All quantities A_i are those of the proposed moves from state o in a given step m .

Thus, rather than only considering the final accepted state used in the subsequent step $m + 1$, we instead include information on all proposed paths out of state o at step m in both the numerator and the denominator. All summands in the numerator and denominator implicitly depend on the current step m in the Markov chain; this dependence is omitted for simplification of notation.

This optimal estimator may be employed for both waste recycling algorithms studied in this paper, MUP and MDP. However, as the results based on multiple displacement proposals (MDP) are much more promising across a range of loadings, we restrict our application of the optimal estimator to simply the MDP algorithm. In these instances, we shall annotate the results as MDP-DJ, for Delmas and Jourdain.

5. RESULTS

The details of the GPU Monte Carlo simulation results are given in this section. The statistical accuracies for different GPU algorithms are compared for various numbers of methane molecules in both unoptimized and optimized CUDA configurations. In the unoptimized CUDA configuration, the number of blocks per SM is set to be 1, and the number of threads per block is set to be 32, resulting in suboptimal work allocation to the GPU threads. In the optimized CUDA configuration, both the number of blocks and the number of threads per block are increased to maximize occupancy in the SM and subsequently performance. Despite its inefficiency, the unoptimized CUDA configuration serves as a good basic starting configuration that allows us to easily test the correctness of the code and to compare the different GPU parallelization methods. Within the optimized CUDA configuration, we compare the performance difference for different distributions of CUDA blocks/threads for the waste recycling Monte Carlo algorithm. We analyze the effect of including the optimal estimator in the statistical accuracy of these WRMC simulations. As a performance comparison, we show speedup numbers of our GPU algorithms compared to a single core CPU code as evidence of performance benefits of using CUDA. Finally, we investigate the performance utility of the waste recycling Monte Carlo algorithm in computing the free-energy profile of methane gas molecules inside a zeolite framework.

5.1. Initial Unoptimized Assessment. Figure 3a displays the single iteration wall time for the embarrassingly parallel Monte Carlo (PMC), monoproposal parallel Lennard-Jones calculation (PLJ), multiple-uniform-proposals waste recycling Monte Carlo (MUP), and multiple-displacement-proposals waste recycling Monte Carlo (MDP), in that order from left to right for $N = 32, 48, 64, 80, 96, 112, \text{ and } 128$ methane molecules. The iteration time for the embarrassingly parallel Monte Carlo method is the longest for all system sizes, which is reasonable given that the number of floating point operations is the largest here since each CUDA thread conducts its own Monte Carlo simulation. The iteration time for the parallel Lennard-Jones calculation is the shortest, as n_{threads} threads share the same methane-MFI system, in contrast to PLJ, and as only one proposal move per CUDA block is required at each iteration as opposed to multiple trial moves per iteration for the two waste recycling methods. In the limiting case of a very large system where Lennard-Jones calculation completely dominates total wall time, the number of floating point operations in the two waste recycling methods would be roughly n_{threads} times larger than the number found in the parallel Lennard-Jones method for the same number of accumulation steps. However, this workload can be reduced by

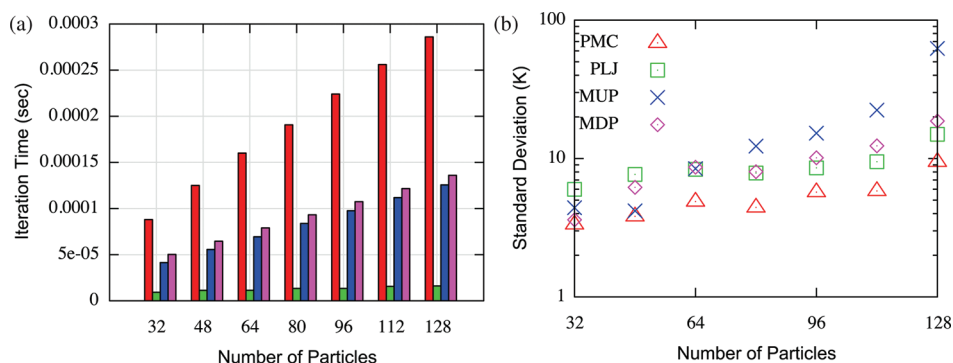


Figure 3. Comparison of the various parallel algorithms for different numbers of methane molecules. (a) Single MC iteration time (in seconds on the left axis) as a function of the number of methane molecules (from left to right: PMC, PLJ, MUP, MDP). (b) Comparison of the standard deviations in the energy (in units of Kelvin) as a function of the number of methane molecules for the various parallelization strategies. The number of accumulation steps is varied to equate the total wall times of the simulations in all four methods.

approximately half in the waste recycling methods as n_{threads} independent proposals all share the same old system energy value at the start of each Monte Carlo step, and thus this value needs to be just calculated once for all of the multiple proposals. Overall, the wall time for the parallel Lennard-Jones method in the limiting case would be approximately $0.5n_{\text{threads}} = 16$ times faster than the waste recycling methods with the WRMC algorithms collecting $32\times$ more energy samples during the simulation. Analyzing the wall times for the two waste recycling methods, we observe that the displacement WRMC has about a $10 \mu\text{s}$ longer iteration time than the uniformly sampled WRMC for all system sizes due to the additional overhead from generating two separate random walks in the displacement WRMC method.

As an initial comparison of the effectiveness of each technique in variance reduction, we run simulations with comparable total wall times as follows. The total number of equilibration and decorrelation steps is set to be 200 000 each while the baseline accumulation step is set to be 1 million for the embarrassingly parallel Monte Carlo method. The accumulation steps for other methods are adjusted with respect to this baseline number such that the total wall time spent inside the GPU for all of the methods remains equal. The number of accumulation steps for parallel Lennard-Jones is set to be 9.5, 11.0, 14.0, 16.7, 16.4, and 17.8 million; the number for uniformly sampled WRMC is set to be 2.1, 2.2, 2.3, 2.3, 2.3, and 2.3 million steps, and the number for displacement WRMC is set to be 1.8, 1.9, 2.0, 2.0, 2.1, and 2.1 million steps for $N = 32, 48, 64, 80, 96, 112,$ and 128 methane molecules. The total number of CUDA blocks, N_{blocks} is equal to 14 with one block occupying each of the streaming multiprocessors. The CUDA block size n_{threads} is set to 32, which leads to a low occupancy number (i.e., $32/1536 = 0.021$, with 1536 being the maximum number of resident threads per multiprocessor in Tesla C2050) and underutilization of the Fermi GPU hardware. Because the main focus here is to investigate the general behavior trends for each of the different CUDA parallelization methods at varying particle densities, we initially explore a simple, reduced layout of thread blocks and sizes and forego optimization analysis until later.

In Figure 3b, we plot on a log-linear scale the standard deviation of total energy for 40 independent Monte Carlo simulation runs as a function of the number of methane molecules for the four methods. Overall, the embarrassingly parallel Monte Carlo (PMC) algorithm provides the best statistical accuracy for all system sizes, as the benefit of processing n_{threads} times more

independent systems than other methods outweighs the cost of the longer single iteration time. In the parallel Lennard-Jones (PLJ) method, the proportional wall time spent in routines other than the Lennard-Jones kernel remains large compared to other methods (due to short single iteration wall time), and this overhead causes performance degradation that is evident especially in the low-density regime where the method fairs the worst. The multiple-uniform-proposals (MUP) WRMC excels for small system sizes/low densities where the likelihood of sampling non-negligible Boltzmann weights (i.e., low energy configurations) in the proposed trial states remains high. However, at large system sizes/high densities, the method fairs the worst, as most of the randomly generated trial states result in high energy configurations that add negligible contribution to the Monte Carlo statistics. This problem is remedied in the multiple-displacement-proposals (MDP) WRMC method where the random walks are conducted in small step sizes and lead to relatively lower energy configurations for each of the proposed trial states. From Figure 3b, we observe that the standard deviation values from the MDP waste-recycling algorithm are comparable to the ones from the PLJ method for all system sizes.

5.2. Exploration of MDP Parameters. Next, we further analyze the displacement WRMC method (MDP) and vary the maximum displacement step size, d_{max} , as well as the total length of the random walks (i.e., n_{threads} , which is the CUDA thread block size) in our MC simulations to determine the optimal parameter settings for the WRMC simulation. Because our interest lies mostly in code optimization for high density systems, we restrict our analysis to the displacement WRMC method for 128 methane molecules, which provides better statistical accuracy at high density compared to the multiple-uniform-proposals WRMC. Figure 4 shows the plot of the standard deviation of the total system energy for 40 independent MC simulations as a function of d_{max} with 1 million accumulation steps for total length of random walks $n_{\text{threads}} = 32, 64, 128, 256,$ and 512 . In all of the simulations, d_{max} is the same value along all three spatial directions, and n_{threads} is set to be a multiple of the warp size in all cases to avoid warp divergence and for simplicity. The total GPU wall time for different n_{threads} are not set to be equal here and accordingly, simulations with larger n_{threads} have longer single iteration times and are expected to provide greater statistical accuracies. Given the GPU hardware limit on register size per streaming multiprocessor (32 kB/SM), the maximum number of threads that can occupy a multiprocessor is 512 in our

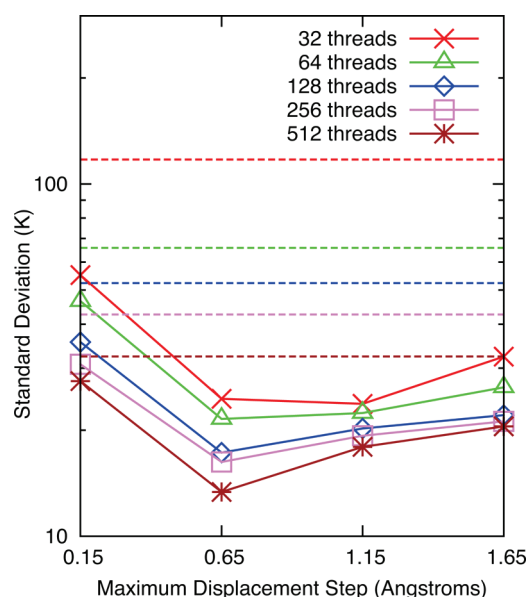


Figure 4. Standard deviation of the average of energy as a function of the maximum individual displacement step, d_{\max} , for CUDA block size ranging from 32 to 512 threads. The CUDA block size corresponds to the total length of the two random walks in the multiple-displacement-proposals (MDP) WRMC algorithm. The standard deviation values for the multiple-uniform-proposals (MUP) WRMC are represented as dashed lines with the same colors (corresponding to same CUDA block size) as the displacement WRMC.

code, thus limiting the length of the random walks. In order to conceive longer random walks, threads from multiple blocks need to be assigned to process the same methane–MFI system, which would effectively replace the one CUDA block per one methane–MFI system mapping that is being currently utilized. In such an implementation, threads from different CUDA blocks would need to be synchronized at each MC step, which can only be achieved through termination of the CUDA kernel given the lack of universal barrier synchronizations in CUDA. Due to the increased latency resulting from relaunching a CUDA kernel at every MC iteration and the diminishing return in statistical accuracy for large thread block sizes, we do not allow for multiple CUDA blocks to process a single system and limit the maximum length of the random walk at 512.

As can be seen from Figure 4, the statistical accuracy in our simulations improves with larger numbers of n_{threads} , as expected since more threads contribute to waste recycling. For all values of n_{threads} , a small d_{\max} (e.g., 0.15 Å) leads to large standard deviation values. While the trial state particle positions near the original position will have reasonably high Boltzmann weights, they are highly correlated with the original position and as such do not add much new information. As d_{\max} is increased for all n_{threads} , the standard deviations begin to increase again. In this case, while the proposals are more decorrelated from the original position, they are far more likely to have a small Boltzmann weight. As discussed below and in more detail in Appendix , we may capture the scaling of these standard deviations with a simple quantitative model, but regardless of the exact scaling, Figure 4 demonstrates that, in all cases presented here, the standard deviation in WRMC using multiple displacement proposals (MDP) substantially improves over multiple uniform proposals (MUP). The displacement WRMC method becomes equivalent

to the uniform WRMC method at d_{\max} values equal to dimensions of the simulation box, a distance which is substantially larger than 1.65 Å.

The basic scaling behaviors of the standard deviation as a function of d_{\max} and n_{threads} may be quite easily understood from the probability of sampling different distances from the generating point. Three considerations are necessary reflecting the purpose of displacement-based Monte Carlo simulation to gather new *relevant* information by optimizing for sampling near the particle's current position, but not too near that position. (1) Sampling regions too close to the original position of the particle ($r < R_{\min}$) does not yield useful additional information, as these positions are within the exclusion zone of the original particle. (2) Sampling regions too far from the original position of the particle ($r > R_{\max}$) also does not yield useful information, as these more dispersed locations are less likely to be in the important regions of phase space. (3) Sampling a local region in space with a higher density of points is no longer useful beyond some density ρ_{cap} . Disregarding the important effect of ρ_{cap} for the moment, for a fixed number of particles, as the step size increases, initially the number of points between R_{\min} and R_{\max} increases, leading to better sampling. Then, the number of points begins to decrease again, leading to somewhat poorer sampling of the important nearby region. For a fixed step size, as the number of proposals increases, the number of samples within that window increases in a *nonlinear* fashion, re-emphasizing why the standard deviation decreases as n_{threads} increases. The effect of ρ_{cap} becomes particularly important for large values of n_{threads} and small d_{\max} , as this combination can lead to a remarkably high number of samples at given radii. Exclusion of point densities exceeding ρ_{cap} is partially responsible for the comparably large standard deviation at small d_{\max} even for large n_{threads} as seen in Figure 4.

A random walker model motivated by these very basic considerations of which regions of space are important to sample is developed in Appendix . Hypothetically, this model could allow for the optimal choice of d_{\max} and n_{threads} to minimize standard deviation for computation time. However, detailed knowledge of molecular organization in the nonuniform environment is required to yield an accurate model. As such, while we can determine an optimal choice of d_{\max} for these systems, in general this is likely not accessible. Therefore, based on the results displayed in Figure 4, we utilize $d_{\max} = 0.65$ Å for the remainder of these studies since that seems reasonably advantageous for all considered n_{threads} and does not rely on optimizing d_{\max} .

5.3. Optimal Estimator of Delmas and Jourdain. We also examine the effect of the optimal estimator on standard deviations. As shown in Figure 5, for both 32 threads and 128 threads, the linear combination of the traditional MC and WRMC results via the factor b^* yields a reduction in the standard deviation of the energy. The estimation of b^* itself does come at a small computational cost; however, the benefits in standard deviation decrease outweigh this computational cost. The optimal estimator appears to have a more beneficial effect for simulations with greater numbers of threads. This likely is a consequence of the improved estimation of b^* itself as more terms are included in the waste-recycling average. Since the use of the optimal estimator provides enhanced accuracy of prediction with minimal computational cost, all further results for multiproposal WRMC shall employ the optimal estimator. The values of b^* are shown in Table 1.

5.4. Optimization of Thread Block Configuration. We now begin to optimize our CUDA configuration. In the previous

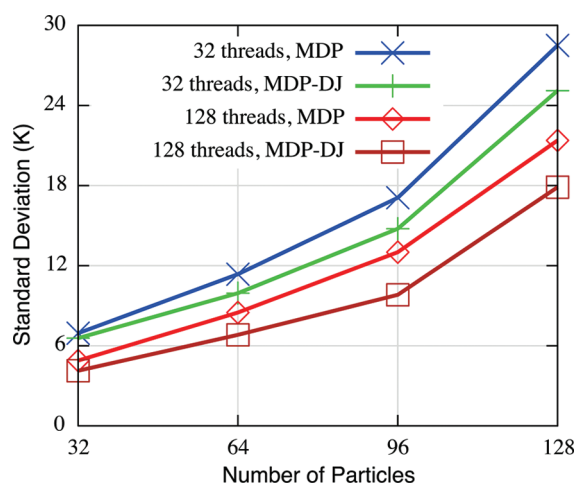


Figure 5. Standard deviation of the average energy as a function of the number of particles for $n_{\text{threads}} = 32$ and 128 for the multiple-displacement-proposals algorithm with (MDP-DJ) and without (MDP) the optimal estimator for 1 million accumulation steps.

Table 1. Values of b^* for MDP-DJ, Which Are Found Taking the Averages of $\langle E \rangle$, $\langle E^2 \rangle$, and $\langle (E_{m+1} - E_m)^2 \rangle$ from the 14 CUDA Blocks, Which Processes Independent MFI Frameworks

n_{threads}	number of particles (N_{tot})			
	32	64	96	128
32	66.74	136.20	224.92	400.40
128	43.14	89.93	162.07	321.72

displacement WRMC simulation results, the total number of CUDA blocks is fixed at 14 with one streaming multiprocessor executing a single CUDA block. This mapping prevents the GPU hardware from scheduling warps from different thread blocks and leads to hardware underutilization, which is especially damaging for small n_{threads} at low occupancy. In order to remedy this situation and to determine the optimal block size/number, we increase the number of CUDA blocks and run simulations with different combinations of CUDA block size (denoted as n_{threads}) and the number of CUDA blocks per multiprocessor (denoted as n_{blocks}). The single iteration wall time numbers for CUDA thread configurations $(n_{\text{threads}}, n_{\text{blocks}}) = (32, 16), (64, 8), (128, 4), (256, 2),$ and $(512, 1)$ are plotted in Figure 6 for $N_{\text{tot}} = 32, 64, 96,$ and 128 methane molecules. In all of the simulations, $n_{\text{blocks}} \times n_{\text{threads}} = 512$, and the total number of CUDA threads is fixed at $512 \times 14 = 7168$. The total number of independent methane-MFI system is $14n_{\text{blocks}}$.

From Figure 6, it can be seen that the single iteration time decreases from the configuration (32, 16) to (64, 8) for all N_{tot} . Because the maximum resident number of thread blocks per multiprocessor is limited to eight in Fermi Tesla C2050 cards, only eight out of the 16 n_{blocks} can concurrently occupy a streaming multiprocessor in (32, 16), resulting in an occupancy number of only $256/1536 = 0.166$ as opposed to $512/1536 = 0.333$ for other configurations, which all have eight or less n_{blocks} . The GPU hardware can be more fully utilized in configuration settings with a higher occupancy number, and subsequently we observe the wall time reduction from (32, 16) to (64, 8).

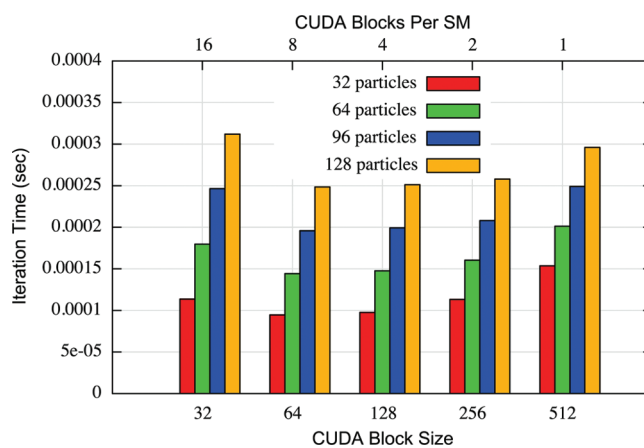


Figure 6. Iteration time as a function of CUDA block size and number of blocks per streaming multiprocessor for 32, 64, 96, and 128 particles (from left to right, respectively) using the MDP-DJ algorithm. The total number of CUDA blocks is $14n_{\text{blocks}}$.

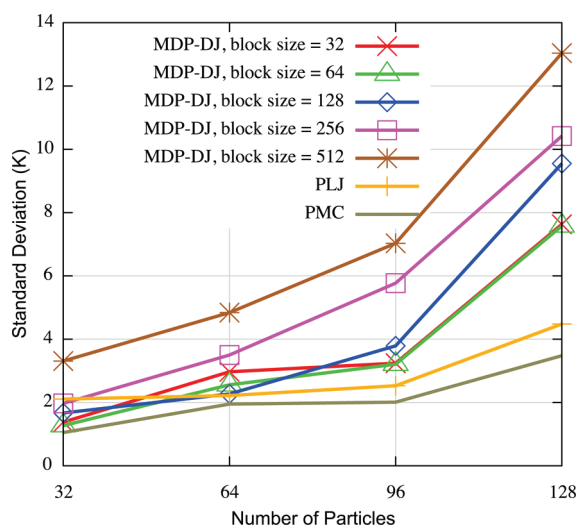
For larger n_{threads} , the wall time increases from (64, 8) as the proportional time spent in the kernel that generates the length n_{threads} random walks becomes larger. Because threads having different n_{blocks} generate the random walks in parallel in our implementation, only the value of n_{threads} largely determines the kernel wall time. Overall, these two factors work together to make the configuration $(n_{\text{threads}}, n_{\text{blocks}}) = (64, 8)$ possess the shortest wall time for all N_{tot} .

It is important to keep in mind that we cannot easily determine the optimal block/thread size by comparing just the iteration wall times since each configuration possesses different values of n_{threads} and provides varying statistical accuracy. In order to meaningfully evaluate the best $(n_{\text{threads}}, n_{\text{blocks}})$ configurations, we follow the same strategy used to derive results in Figure 3b and set the accumulation steps for (32, 16) to be a baseline number of 1 million and adjust the number of steps in other configurations to equate the total wall time spent in the GPU for all $(n_{\text{threads}}, n_{\text{blocks}})$. Again, 40 independent displacement WRMC simulation runs that include the optimal estimator are conducted, and the average energy and standard deviation value are tabulated in Table 2 (the values for n_{blocks} are the same ones used in Figure 6 and omitted in the labels). For comparison, we include simulation results from parallel Lennard-Jones and embarrassingly parallel MC methods with equal wall time for all N_{tot} and all of the results are plotted in Figure 7 to illustrate the behavior. In the parallel Lennard-Jones method, the (32, 16) configuration was utilized for all N_{tot} , as this results in minimum single iteration wall time. In the embarrassingly parallel MC method, we used the (64, 8) configuration, which also provided the minimum single iteration wall time. For all N_{tot} the lowest standard deviation values among the displacement WRMC methods are observed for the (32, 16) and the (64, 8) configurations, as the performance cost of reducing the number of independent ensembles outweighs the benefit of increasing the number of WRMC proposal trial states for $n_{\text{threads}} > 64$. Overall, similar to results found from the nonoptimized configurations in Figure 3b, the overall most accurate results were found from the embarrassingly parallel MC algorithm.

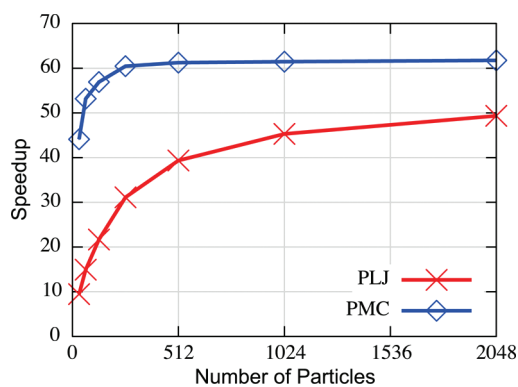
5.5. Comparison of GPU to CPU Timing. We also compare the efficiency of the GPU methods to that of a single core CPU monoproposal MCMC method. The algorithm for CPU MCMC

Table 2. Average Energy and Standard Deviation Values for 32, 64, 96, and 128 Particles for Different n_{threads} in Units of Kelvin for the MDP-DJ Algorithm

n_{threads}	number of particles (N_{tot})			
	32	64	96	128
32	-68354.59 ± 1.39	-138027.62 ± 2.97	-209278.78 ± 3.24	-281343.57 ± 7.63
64	-68353.49 ± 1.27	-138027.38 ± 2.56	-209278.40 ± 3.21	-281340.17 ± 7.58
128	-68354.19 ± 1.67	-138027.66 ± 2.27	-209280.23 ± 3.73	-281340.30 ± 9.55
256	-68355.18 ± 1.97	-138028.18 ± 3.50	-209278.08 ± 5.77	-281343.87 ± 10.42
512	-68352.43 ± 3.32	-138028.45 ± 4.84	-209277.23 ± 7.04	-281344.35 ± 13.05

**Figure 7.** Standard deviation in energy for the multiple-displacement-proposals (MDP-DJ) WRMC results from Table 2 as well as parallel Lennard-Jones (PLJ) and embarrassingly parallel simulations (PMC). The block sizes of 32 and 64 (i.e., $n_{\text{threads}} = 32$ and 64) provide the lowest standard deviation for the MDP-DJ algorithm.

is similar to the one used in the parallel Lennard-Jones method except for the serial processing of the pair-potential calculation. The CPU MCMC source code is compiled using an Intel 11.1 compiler, which provides faster CPU single iteration time compared to gcc 4.4.2 for small system sizes. We define speedup as the ratio between the CPU and the GPU single iteration times and plot the results in Figure 8. It is difficult to compare the single iteration wall times with the WRMC algorithm and arrive at any meaningful conclusion since the two methods collect samples with uneven importance weights. Accordingly, we choose to omit this comparison and concentrate on GPU results from the parallel Lennard-Jones and the embarrassingly parallel methods. Unlike previous simulations, we simulate relatively large system sizes that provide unphysical high loading situations to better assess the speedup behaviors. Accordingly, the energy results are erroneous for large system sizes, which is acceptable in this context since only the numerical speedup results are meaningful and of interest here. As can be seen from Figure 8, the embarrassingly parallel MC method has better performance over the parallel Lennard-Jones method for all system sizes with a maximum speedup value of 61.75 at 2048 methane molecules. However, for even larger system size, we expect the two methods to provide similar values of speedup as the overhead from routines involved in tasks other than pair-potential calculation become

**Figure 8.** Speedup as a function of the number of particles for system sizes from 32 to 2048 methane molecules for double-precision (DP) calculations using the Fermi Tesla C2050 card. The speedup is defined as a ratio between single iteration CPU and a single iteration GPU wall time. The maximum speedup of 61.75 is observed for 2048 methane molecules.

negligible for the parallel Lennard-Jones method, which is the cause of slower speedup in small systems. For particle loadings of interest in the methane–MFI system, at 32, 64, and 128 particles, there are respectively 9.47 (44.11), 14.93 (53.18), and 21.68 (56.90) speedups in the parallel Lennard-Jones (embarrassingly parallel) method over the CPU results.

5.6. Application to Free Energy Calculation. Finally, we consider a simulation scenario where waste recycling was shown to be useful in the literature—the calculation of the free-energy profile along a reaction coordinate. The forms of $P(z)$ and $\beta F(z)$ shown in Figure 9 nicely illustrate the simple free-energy barrier in the windows between two cages, with mild corrugation within the cages due to the packing of those methane molecules occupying the cages.

In the calculation of these histograms $P(z)$, the previous efforts in this paper toward optimization on the GPU are portable, so we determine $P(z)$ simply using the GPU-optimized mappings. *A priori*, we expect this to be a scenario where waste recycling Monte Carlo could lead to faster convergence in the histogram because the higher free-energy states within the window between cages will be sampled by the Markov chain of states much less frequently, yet some subset of the multiple proposals not accepted likely probes this region of space. And indeed, our expectations are borne out. For simulations of 32 methane molecules in the zeolite LTA for 10 million MC cycles, $P(z)$ is calculated via simple binning. In this instance, both parallel Lennard-Jones (PLJ, not shown) and waste-recycling with multiple displacement proposals (MDP, Figure 9) give quite similar forms for $P(z)$.

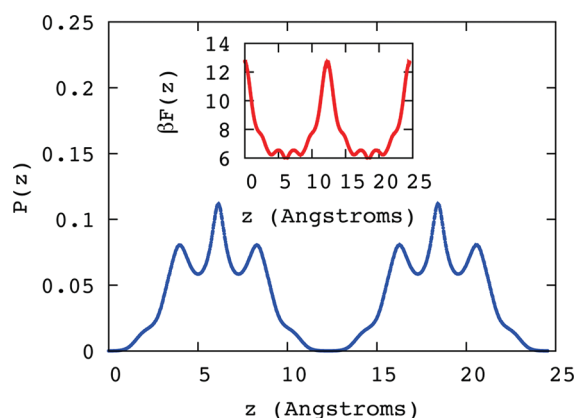


Figure 9. Plots for $P(z)$ vs z (blue) and $\beta F(z)$ vs z (red, inset) for LTA with 32 methane molecules (10 million Monte Carlo cycles using the waste recycling with multiple displacement proposals (MDP) with $n_{\text{threads}} = 32$ and $n_{\text{blocks}} = 8$). $P(z)$ indicates the histogram of probabilities, while $\beta F(z)$ represents the free-energy (unitless) along the z direction of the LTA (unit cell length = 24.555 Å). The free-energy barrier is graphically represented by the bump near $z = 0.5 \times 24.555$ Å in the inset curve.

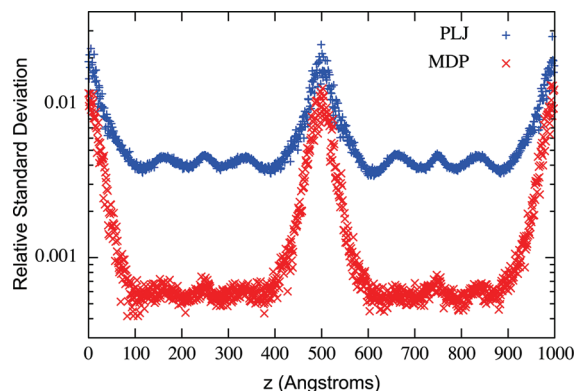


Figure 10. Relative standard deviation as a function of position across the LTA zeolite for PLJ and MDP simulations with equal wall times.

To assess timing gains, we run simulations using both techniques such that the total wall time is fixed (10 million MDP cycles and 51.2 million PLJ cycles). For the MDP simulations, we exclude the optimal estimator of Delmas and Jourdain as the system size (e.g., 32 methane molecules) is sufficiently small that the overhead required in obtaining relevant quantities would most likely not decrease the standard deviation taking into account wall time. Moreover, the main point here is to emphasize the performance difference between the waste recycling method and other more conventional parallel methods. We consider standard deviations across a set of 20 MC simulations for each parallelization strategy. In order to better quantify any computational gains, we consider a single point in these normalized probability distributions, $P(z^*)$. We choose z^* to be the least probable location, with z^* thereby corresponding to the free energy maximum, and $P(z^*)$ is a quantity important in determining diffusion rates through zeolites.¹⁷ For the system of 32 methane molecules in LTA, parallel LJ yields $2.41 \times 10^{-4} \pm 4.04 \times 10^{-6}$, while waste recycling Monte Carlo with multiple displacement proposals yields $2.44 \times 10^{-4} \pm 1.13 \times 10^{-6}$. In Figure 10, we show the relative standard deviation (i.e., standard deviation divided by the mean value) at all of the bins across a single dimension of the

LTA zeolite for the same sets of simulations. As can be seen, the MDP outperforms PLJ across the entire set of bins. Thus, as hypothesized, waste recycling Monte Carlo on a GPU is computationally advantageous, oversimplifying accelerating the Lennard-Jones potential calculation when sampling *rare* events.

6. CONCLUSIONS

Waste recycling Monte Carlo has previously been proposed as a way to leverage the “unused” rejected trial states in Monte Carlo simulations. Two examples of scenarios where waste recycling has proven particularly useful are (1) parallel tempering where there exists a large amount of follow data available for immediate harvesting and (2) calculation of potentials of mean force which rely on histogram collection.^{8,17,18}

Here, we examine in detail the application of WRMC to the calculation of the simple average of energy E in canonical Monte Carlo simulations, where waste recycling is not particularly beneficial for single trial moves. Such an approach has allowed us to optimize the GPU implementation of several distinct parallelization strategies. The strategies are (1) embarrassingly parallel Monte Carlo (PMC), (2) parallel Lennard-Jones calculation (PLJ), (3) waste recycling based on multiple uniform proposals (MUP), and (4) waste recycling based on multiple displacement proposals (MDP). Our figure of merit in these studies of $\langle E \rangle$ has been the uncertainty in this average for each technique, with a fixed computational wall time.

Our analysis of various GPU-implementation optimizations for calculating the average energy has found that the use of parallel Monte Carlo simulations is the most computationally efficient approach. Such a conclusion is not surprising because the sampling of n_{threads} -fold more independent simulation systems than for PLJ or WRMC leads to a greater reduction in variance. However, simply by virtue of limited memory resources, such an embarrassingly parallel approach to MC simulation will not always be feasible.

Allocating one simulation to an entire thread block eases those memory constraints while leading to fewer total simulations. We examined two possible approaches for using the computational resources within a thread block—parallelization of the Lennard-Jones pair potential calculation and the evaluation and waste recycling of multiple Monte Carlo proposals. Among these techniques, PLJ is the most efficient computational approach for evaluating the average energy, allowing for the fastest propagation through phase space based on CPU wall time. In essence, this again is simply because efficient propagation through phase space via the PLJ algorithm dominates for variance reduction in the average energy.

The effectiveness of the multiple-uniform-proposals WRMC algorithm decreases with greater density of particles as the trial states generated from the uniformly random distribution provide high energy states that make negligible contributions to the statistics. Generating trial states based on displacements, as done in the multiple-displacement-proposals WRMC algorithm, side-steps this problem. With careful generation of displacement trial states based on a random walk of displacements, very little modification of the remainder of the WRMC algorithm is required.

When calculating simple MC averages such as $\langle E \rangle$, parallel Monte Carlo simulations with one simulation per thread (PMC) is the algorithm of choice, followed by parallelization of the Lennard-Jones calculations (PLJ) within a thread block. Neither WRMC algorithm is as computationally efficient. However, as

noted earlier, in a variety of scenarios related to determination of histograms and sampling of rare events, WRMC can be advantageous on the CPU even when it is not for other quantities.

In particular, the free energy profile of particles in the caged zeolite structures, crucial for characterizing diffusive behavior, is straightforwardly related to the density profile of methane molecules in the zeolite LTA. Therefore, we also compare the optimized version of each of these algorithms for determination of the free-energy profile between adsorption cages within the zeolite LTA, in order to assess if WRMC on the GPU is then worthwhile. Indeed, WRMC with multiple displacement proposals (MDP) is more computationally efficient than either PMC or PLJ in conducting free energy calculations.

Beyond assessing various Monte Carlo algorithms for framework-adsorbate systems, we have also examined the utility of the optimal estimator of Delmas and Jourdain¹⁵ in a multiproposal scenario for the first time. We find that this optimal estimator is advantageous for energy calculations with relatively little computational overhead and provides meaningful improvement in the estimation of properties.

While our studies indicate that often multiproposal waste-recycling algorithms are not computationally the ideal path for employing GPU resources in MC simulations, waste recycling can be beneficial over other parallelization strategies for sampling rare events. Our findings suggest that the relative merits of WRMC in a parallel GPU computing environment are comparable to those in a single CPU computing environment, with waste-recycling Monte Carlo advantageous for sampling rare events but less useful for straightforward MC averages. However, the careful implementation of a multiproposal framework as developed in this paper is necessary to even employ waste recycling in the GPU environment. Furthermore, we are currently expanding the work on the other GPU implementations (PMC and PLJ) in order to substantially accelerate calculation of adsorption isotherms in zeolites, as this is necessary to computationally screen the millions of hypothetical zeolite structures.²³

APPENDIX A. COMPILER AND PROCESSOR DESCRIPTION

The NERSC cluster used, Dirac, is a testbed GPU cluster that contains 44 Fermi Tesla C2050 GPUs, which come equipped with 448 CUDA cores, 14 streaming multiprocessors (SMs), 3 GB of DRAM, and ECC memory. The card delivers peak single-precision (double-precision) performance of 1.03 TFlops (515 GFlops) and 144 GB/s of peak memory bandwidth. The PCI Express 2.0 with 16 lanes is used to transfer data back and forth from the CPU to the GPU memory. The CPU node within Dirac consists of two Intel 5530 2.4 GHz, quad core Nehalems with an 8 MB cache, 5.86 GT/s QPI, and 24 GB DDR3-1066 Reg ECC memory. The CUDA compiler driver NVCC along with gcc 4.4.2 with -O3 optimization flag is used for all of the GPU simulations, whereas for CPU simulations, the Intel C++ Compiler 11.1 is used. CUDA Toolkit 3.2 along with CUDA C runtime is used, and the CURAND library is utilized to generate pseudorandom numbers based off of the XORWOW algorithm.¹⁹ The random numbers are generated directly inside the device kernel using the device API, thereby bypassing the need to transfer the numbers from the CPU to the GPU. Random generator state initialization (curand_init()) and random number generation (curand()) are divided into two separate kernels in order to maximize performance. All of the results reported in this work use double-precision (64-bit) floating point numbers.

APPENDIX B. INCORRECT OR INEFFICIENT MULTIPLE DISPLACEMENT PROPOSALS WITH WASTE-RECYCLING

In section 4.4, we described our chosen MDP algorithm with an emphasis placed on the generation of trial positions such that we may use the simple symmetric multiproposal Barker-like acceptance ratio. Here, we describe alternative routes that proved either incorrect or inefficient.

Incorrect MDP Algorithms. A naive, and incorrect, modification of the multiproposal waste recycling algorithm in order to generate multiple displacement proposals would involve generating the proposed particle positions all as single uniform displacements from the original position of particle k . Such an approach yields incorrect results. This stems from the fact that not all points within the set $P = \{o, \{n\}\}$ are accessible to each other via this trial-state generation algorithm. Instead, generating $\{n\}$ based on drawing displacements from a random Gaussian distribution removes this formal accessibility barrier since there is always some finite possibility of any point in A generating the remaining points. However, this set P still has inherent bias towards the state o as the generating point. Delmas and Jourdain¹⁵ define a multiproposal acceptance ratio accounting for the *a priori* probabilities of trial generation based on the associated Gaussian probability density that corrects this state generation bias. However, for the range of numbers of proposals explored in this paper, calculated properties still exhibited a dependence on the number of proposals.

Inefficient MDP with Configurational Bias. The crux of our employed multiproposal algorithms lies in constructing a set of proposals that are equally likely to have proposed that set, and this results in a simply symmetric Boltzmann acceptance ratio. However, a more traditional approach to multiproposal Monte Carlo would be the implementation of configurational bias Monte Carlo (CBMC).⁷ A standard CBMC algorithm for multiple proposals might propose a set of trial states $\{n\}$ based on displacements from the old state o . A proposed state is then chosen solely from the set of trial states $\{n\}$ based on their relative Boltzmann weights. Once a proposed state is chosen, it is either accepted or rejected on the basis of both the sum of the Boltzmann weights from the set of trial states $\{n\}$ and a sum of Boltzmann weights due to a hypothetical set of trial states $\{o'\}$ which are generated by displacements from the proposed state. This generation of forward and backward sets of trial states is required in order to fulfill the condition of superdetailed balance.⁷

Simulating with multiproposal configurational bias yields correct results; however, optimal implementation of the waste recycling approach would require the accumulation of the following summation:

$$A = \sum_{i \in \{n\}} \frac{w_i}{W} \{A_i \text{acc}(o \rightarrow i) + A_o [1 - \text{acc}(o \rightarrow i)]\} \quad (9)$$

where w_i is defined again as $\exp(-\beta E_i)$ and W is defined as

$$W = \sum_{i \in \{n\}} w_i \quad (10)$$

with the old state *excluded* from the summation. Aside from this exclusion, eq 10 is similar to weightings employed thus far for waste recycling.

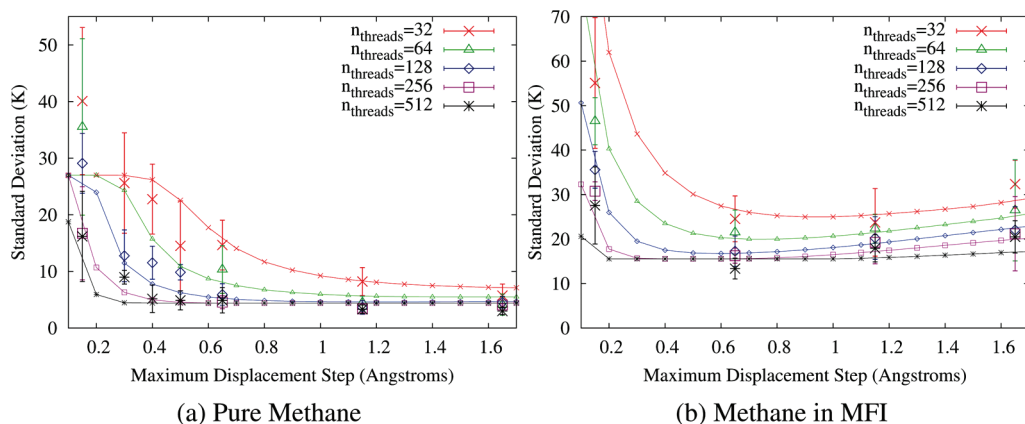


Figure 11. Random walker model of standard deviation in calculated energies for both 128 methane molecules as a uniform fluid (a) and 128 methane molecules in the zeolite MFI (b). Each is modeled via the proportionality in eq 12 with a single proportionality constant A for all n_{threads} and d_{max} . For the uniform methane fluid, $A = 27.0$, $\rho_{\text{cap}}\sigma^3 = 1$, $R_{\text{min}} = 3.8 \text{ \AA}$, and $R_{\text{max}} = 8.0 \text{ \AA}$. For methane adsorbed in MFI, $A = 110.0$, $\rho_{\text{cap}}\sigma^3 = 10$, $R_{\text{min}} = 1.0 \text{ \AA}$, and $R_{\text{max}} = 4.0 \text{ \AA}$. Choice of these parameters is discussed in the text. For each plot, the data from GPU calculations are displayed using the larger symbols with error bars, and the results from the simple random walker model are presented with the smaller symbols and lines to guide the eye.

The inefficiency in implementing waste recycling for configurational bias lies in the definition of acceptance probabilities. In order to determine an acceptance ratio for a specific trial state i , we must generate a set of trial old states $\{o'|i\}$ from the state i in order to obey superdetailed balance. Given these *unique* sets of old trial states, the acceptance probability for any state i is defined as

$$\text{acc}(o \rightarrow i) = \min \left[1, \frac{W}{W(\{o'|i\})} \right] \quad (11)$$

Since a different set $\{o'|i\}$ must be generated for each proposed trial state i , a total of $N_{\text{prop}}(N_{\text{prop}} - 1)$ new particle positions and new energies must be evaluated in order to fully leverage waste recycling coupled to multiproposal MC, instead of N_{prop} energies, as is the case for our MDP algorithm. A waste recycling expression similar to the above CBMC algorithm has been successfully and efficiently employed by Athènes and Calvo¹⁸ in the context of replica exchange. However, in replica exchange simulations, the associated acceptance ratios require no new calculations of energies and simply require the rescaling of previously determined energies by new factors β .

Our proposed approach for multiproposal waste recycling involving the construction of random walks, as stated in subsection 4.3, yields correct averaged results which are invariant to the number of proposals. Furthermore, it allows inclusion of all calculated energies into the waste-recycling expression for accumulating averages. In the implementation based on constructing a random walk chain, no calculated energies lie fallow. As an added benefit, since this approach is constructed via Barker-like acceptance ratios rather than Metropolis-like acceptance ratios, we may employ the optimal estimator discussed in section 4.5.

APPENDIX C. SIMPLE MODEL FOR MDP PARAMETERS

The variance in the average energy for the MDP waste-recycling algorithm has a rather unusual functional form, as shown in Figure 4. With the considerations outlined in section 5.2, we may quantitatively model the variation in standard deviation as a function of d_{max} and n_{threads} . We display our model results in Figure 11 for both the standard deviations for methane in MFI and for pure methane. For each value of d_{max} and n_{threads} ,

we determine the number densities associated with finding a position on the random walk a certain distance r from the origin. These densities are calculated on the basis of a total of 10^6 distinct random walks constructed numerically following the algorithm employed in this paper for the WRMC simulations and accounting for periodicity as the paths are constructed. Each number density $N(r; d_{\text{max}}, n_{\text{threads}})$ is initially calculated such that $\int_0^\infty N(r; d_{\text{max}}, n_{\text{threads}}) dr = n_{\text{threads}} - 1$ and is subsequently capped at each r to be at maximum $\rho_{\text{cap}} 4\pi r^2$. This $N(r; d_{\text{max}}, n_{\text{threads}}, \rho_{\text{cap}})$ thereby encompasses effects due to the spherical geometry at each r as well as the limits on the effectiveness of higher sampling number density. The standard deviation on a calculation is hypothesized to scale inversely with the square root of the number of meaningfully sampled points. We express this as

$$\text{std.dev.} \propto \frac{1}{\sqrt{1 + \int_{R_{\text{min}}}^{R_{\text{max}}} N(r; d_{\text{max}}, n_{\text{threads}}, \rho_{\text{cap}}) dr}} \quad (12)$$

where the original position is always included as a meaningful point and all other sampled points are solely included if they are between R_{min} and R_{max} and have not exceeded the local density ρ_{cap} .

As shown, in Figure 11, this basic model captures semiquantitatively the features of the standard deviation of energy as a function of d_{max} and of n_{threads} for both a uniform Lennard-Jones fluid of 128 methane molecules with the MFI framework removed (corresponding to $\rho\sigma^3 = 0.156$) and for 128 methane molecules adsorbed in eight unit cells of the zeolite MFI. For methane as a uniform LJ fluid, ρ_{cap} is set by the relationship $\rho_{\text{cap}}\sigma^3 = 1$, and R_{min} and R_{max} are set to 3.8 \AA and 8.0 \AA , approximately σ and 2σ , perfectly reasonable parameters for a moderately dense gas. For methane adsorbed in the zeolite MFI, we used substantially different parameters based on knowledge of the organization of methane in MFI. For this degree of adsorption, the methane sites are separated by approximately 5 \AA , and within the channels, they organize in a single file. As such, we set R_{min} to be 1.0 \AA in order to account for meaningful sampling within the radius of the MFI channel. And we choose R_{max} to be 4.0 \AA since distances greater than this yield configurations overlapping with their nearest neighbors. Given that the highly

packed and nonuniform nature of this system introduces greater corrugations, we choose ρ_{cap} to be 10 times higher. The agreement between our simple random walker model and the GPU calculated standard deviations is quite favorable for both the uniform and nonuniform systems, as displayed in Figure 11.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jihankim@lbl.gov; Berend-Smit@berkeley.edu.

ACKNOWLEDGMENT

We thank Joseph Swisher and Mahmoud Forrest Abouelnasr for useful discussion and guidance in simulating zeolites. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. B.S. was supported as part of the Center for Gas Separations Relevant to Clean Energy Technologies, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number DE-SC0001015. J.K. was supported by the Director, Office of Science, Advanced Scientific Computing Research, of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. J.M.R. acknowledges the support of the Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy, FWP number SISGRKN.

REFERENCES

- (1) Frenkel, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17571–17575.
- (2) NVIDIA CUDA Programming Guide 3.0. http://developer.download.nvidia.com/compute/cuda/3_0/toolkit/docs/NVIDIA_CUDA_ProgrammingGuide.pdf (accessed August 29, 2011).
- (3) (a) Anderson, J.; Lorenz, C.; Travesset, A. *J. Comput. Phys.* **2008**, *227*, 5342–5359. (b) Stone, J.; Phillips, J.; Freddolino, P.; Hardy, D.; Trabuco, L.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
- (4) Preis, T.; Virnau, P.; Paul, W.; Schneider, J. *J. Comput. Phys.* **2009**, *228*, 4468–4477.
- (5) Li, H.; Petzold, L. *Int. J. High Perform. Comput. Appl.* **2010**, *24*, 107–116.
- (6) Anderson, A.; Goddard, W.; Schroder, P. *Comput. Phys. Commun.* **2007**, *177*, 298–306.
- (7) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, CA, 2002.
- (8) Athènes, M.; Marinica, M.-C. *J. Comput. Phys.* **2010**, *229*, 7129–7146.
- (9) Smit, B.; Maesen, T. L. M. *Chem. Rev.* **2008**, *108*, 4125–4184.
- (10) Smit, B. *J. Phys. Chem.* **1995**, *99*, 5597–5603.
- (11) Dubbeldam, D.; Calero, S.; Vlugt, T.; Krishna, R.; Maesen, T.; Beerdsen, E.; Smit, B. *Phys. Rev. Lett.* **2004**, *93*, 088302.
- (12) Dubbeldam, D.; Calero, S.; Maesen, T.; Smit, B. *Phys. Rev. Lett.* **2003**, *90*, 245901.
- (13) Dubbeldam, D.; Beerdsen, E.; Vlugt, T.; Smit, B. *J. Chem. Phys.* **2005**, *122*, 224712.
- (14) Dubbeldam, D.; Calero, S.; Vlugt, T.; Krishna, R.; Maesen, T.; Smit, B. *J. Phys. Chem. B* **2004**, *108*, 12301–12313.
- (15) Delmas, J.-F.; Jourdain, B. *J. Appl. Probab.* **2009**, *46*, 938–959.
- (16) Beerdsen, E.; Dubbeldam, D.; Smit, B. *J. Phys. Chem. B* **2006**, *110*, 22754–22772.
- (17) Coluzza, I.; Frenkel, D. *ChemPhysChem* **2005**, *6*, 1779–1783.
- (18) Athènes, M.; Calvo, F. *ChemPhysChem* **2008**, *9*, 2332–2339.
- (19) CUDA CURAND Library. http://developer.download.nvidia.com/compute/cuda/3_2/toolkit/docs/CURAND_Library.pdf (accessed August 29, 2011).
- (20) Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (21) NVIDIA CUDA C SDK. <http://developer.nvidia.com/cuda-toolkit-sdk> (accessed August 29, 2011).
- (22) Adjanor, G.; Athènes, M.; Rodgers, J. M. *J. Chem. Phys.* **2011**, *135*, 044127.
- (23) Earl, D.; Deem, M. *Ind. Eng. Chem.* **2006**, *54*, 5449–5454.

Excited States and Absorption Spectra of UF_6 : A RASPT2 Theoretical Study with Spin–Orbit Coupling

Fan Wei, Guo-Shi Wu, W. H. Eugen Schwarz,[†] and Jun Li*

Department of Chemistry and Laboratory of Organic Optoelectronics and Molecular Engineering of the Ministry of Education, Tsinghua University, Beijing 100084, China

S Supporting Information

ABSTRACT: Uranium hexafluoride (UF_6) is an important compound in nuclear chemistry. The theoretical investigation of its excited states is difficult due to the large number of uranium valence orbitals and ligand lone pairs. We report here a detailed relativistic quantum chemical investigation of its excited states up to about 10 eV using restricted active space second-order perturbation theory (RASPT2). Scalar and spin–orbit (SO) relativistic effects are treated by a relativistic small-core pseudopotential. The RASPT2/SO results remain moderately accurate when the electrons in the active space are restricted to single and double excitations. All eight major spectral peaks corresponding to ligand-to-metal charge transfer have been reproduced within an accuracy of about 0.2 eV and are tentatively assigned. We find that BLYP-based hybrid density functional with 35% Hartree–Fock exchange well reproduce the excitation energies of UF_6 .

INTRODUCTION

Uranium hexafluoride (UF_6) is widely used in the uranium enrichment process generating fuel for nuclear reactors and nuclear weapons. Understanding its ground- and excited-state properties is important. Accurate theoretical calculations of molecules containing actinides have been challenging, especially when involving excited states. The difficulties are due to complicated electron correlation effects, to significant scalar relativistic (SR) and spin–orbit coupling (SOC) effects, to the large number of valence and semicore electrons in actinide atoms, and to the complexity of electronic configurations owing to f-electrons. A number of theoretical investigations has been carried out on the ground-state properties of UF_6 , including geometries, electronic structures and spectroscopic properties, using ab initio wave function theory (WFT) and density functional theory (DFT).^{1–8}

For excited states, WFT electron correlation methods based on the multiconfiguration self-consistent field (MCSCF) approach are often needed. In particular, the complete active space self-consistent field approach with second-order perturbation theory (CASPT2) is commonly used to account for both dynamic and nondynamic electron correlations, respectively.^{9,10} CASPT2 with the state functions subsequently interacting through SOC yields reliable transition energies for the low-lying excited states of isolated UO_2^{2+} and $[\text{UO}_2\text{Cl}_4]^{2-}$.¹¹ Early multireference configuration interaction calculations with single and double substitutions (MR-CISD) using relativistic effective core potentials (RECP) performed less satisfactory,^{12,13} mainly due to the large atomic core chosen for uranium. One may expect that RECP calculations with an appropriate “small core” (i.e., with 60 electrons in the frozen uranium core, 32 electrons in the 5s-5p-5d-6s-6p-5f-6d-7s valence shells being optimized) will perform comparable to all-electron calculations, such as those simulating relativistic scalar and SOC effects by the Douglas–Kroll–Hess (DKH) approximation.¹⁴

The comparison of theoretically computed electronic spectra with corresponding experimental data helps to calibrate the

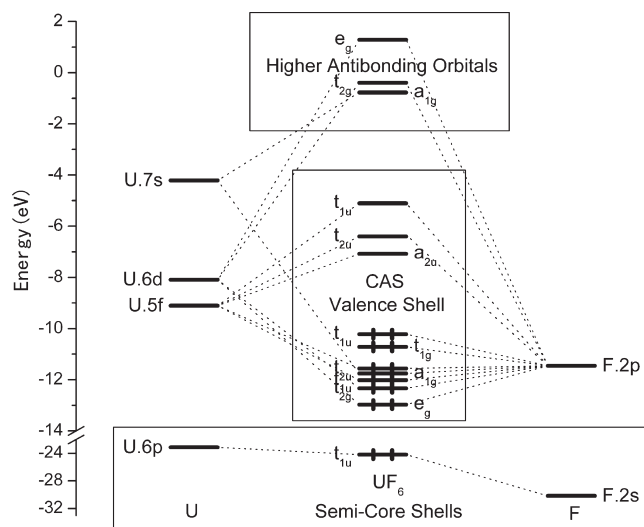


Figure 1. SR Kohn–Sham orbital energy levels of neutral U, F, and UF_6 , in eV.^{3,22}

theoretical methodology and conversely to understand and assign the experimental data. The reported photoabsorption and electron-impact spectra of gaseous UF_6 provide important information about the low-lying excited states.^{15,16} Some of these states involve electronic ligand-to-metal charge transfer (LMCT) and have been investigated theoretically.^{17–22} Already in 1983, Hay²¹ had carried out CIS and CISD calculations on the low-lying excited singlet and triplet states with SOC simulated by a large-core pseudopotential. The excitations correspond to one-electron transitions from the 18 highest occupied orbitals with 36

Received: January 9, 2011

Published: August 10, 2011

electrons, representing the $6F^{-}2p^6$ shells slightly mixed with U-5f,6p,6d,7s atomic orbitals (AOs), to the 7 lowest virtual ones of U-5f type, see the molecular orbital (MO) level scheme in Figure 1.^{21,22} The calculations were quite satisfactory, concerning the low-energy ranges of the experimental spectra. On the other hand, limited by the method, the RECP, the basis sets, and the computer resources of the time, the calculated transition energies of various dipole-allowed ${}^1T_{1u}$ states were less accurate. In particular, the calculated excitation energies above 7 eV deviated from the experimental ones by as much as 0.8–1.5 eV.

To obtain more accurate results, it is necessary to employ more developed post-HF approaches. In principle, complete active space self-consistent field (CASSCF) with an adequate choice of the active orbitals for the nondynamical correlations, CASPT2 energy improvement for dynamical correlation and subsequent SO-CI for the spin–orbit splittings should produce reliable data. The large near-degenerate valence shell of UF_6 easily leads to configuration spaces of many billions of configuration state functions (CSFs), so that excitation restrictions within the active space become mandatory. The restricted active space (RAS) procedure RASSCF/RASPT2²³ has been successfully applied to CuO_2 , Cu_2O_2 , and other systems,^{24,25} such as UO_2 , which shows a similar excitation pattern as UF_6 .^{26,27}

In the present work we have applied the RASSCF/RASPT2/SO approach to various excited states of the UF_6 molecule. The important SOC effect is accounted for by the common RECP procedure,²⁸ where the 1σ containing one-electron operators²⁹ also simulate the two-electron spin-coupling effects. After numerous test calculations we have found that RASSCF/RASPT2/SO approach with RASSCF limited to single and double excitations already provides reasonably accurate results for excitations of the $F-2p \rightarrow U-5f$ type. The experimental excitation energies of UF_6 ¹⁶ within 10 eV have been well reproduced within about 0.2 eV.

Time-dependent density functional theory (TDDFT) often provides an alternative to account for dynamic electron correlation of excited states.^{3,30–32} Previous calculations on actinide complexes indicated that TDDFT excitation energies are of diverse accuracies when comparing with advanced multireference WFT results.^{33,34} Therefore a series of TDDFT calculations has been carried out with various exchange–correlation (XC) functionals to further test the reliability of this method for describing the excited states of actinide systems, such as UF_6 .

METHODOLOGY AND COMPUTATIONAL DETAILS

All ab initio calculations (HF and post-HF) based on WFT were performed with MOLPRO 2008.³⁵ The Stuttgart “small core” RECP and the SDD valence basis [12s11p10d8f]/[8s7p6d4f] were used for the U atom,¹³ while all-electron aug-cc-pVDZ basis sets were applied for the F atoms.³⁶ In the present study, we confine ourselves to vertical excitations. The experimental ground-state geometry of UF_6 (O_h symmetry) with an equilibrium bond length of $R_{U-F} = 1.999 \text{ \AA}$ was taken for all states.³⁷

In the CASSCF approach, a zeroth-order multireference wave function is generated by a full CI procedure within the CAS specified by (N,M) , where N and M are the numbers of correlated electrons and active orbitals. The total number of CSFs rapidly grows with increasing N and M , which virtually limits the applications of CASSCF and CASPT2 to relatively small values of N and M even for present-day computer hardware and software. In the case of UF_6 , however, energy and occupation

Table 1. Irreducible Representations of Low-Lying Single-Electron Excited States of UF_6 ^a

orbital transition	excited states	with triplet SOC
$a_1 \rightarrow a_2$	A_2	T_2
$e \rightarrow a_2$	E	T_1, T_2
$a_1 \rightarrow t_1$ or $t_2 \rightarrow a_2$	T_1	A_1, E, T_1, T_2
$a_1 \rightarrow t_2$ or $t_1 \rightarrow a_2$	T_2	A_2, E, T_1, T_2
$e \rightarrow t_1$ or $e \rightarrow t_2$	T_1, T_2	$A_1, A_2, 2E, 2T_1, 2T_2$
$t_1 \rightarrow t_1$ or $t_2 \rightarrow t_2$	A_1, E, T_1, T_2	$A_1, A_2, 2E, 4T_1, 3T_2$
$t_1 \rightarrow t_2$ or $t_2 \rightarrow t_1$	A_2, E, T_1, T_2	$A_1, A_2, 2E, 3T_1, 4T_2$

^aWithout SOC or with singlet SOC, last column with triplet SOC, and the trivial g and u specifications being omitted.

analysis of the MOs, as discussed in the next section, indicates that the effective active space should comprise at least the above-mentioned 36 $F^{-}2p$ valence electrons and the respective 18 highest occupied and 7 lowest unoccupied orbitals of U-5f type, consistent with the selection of Hay.²¹ Such a selection of the minimum active space results in a total of 25 orbitals with 36 electrons, which still excludes the F-2s and U-6s-6p semicore shells and next higher ones of U-6d and -7s type (see Figure 1). A CASSCF calculations using an active space of (36,25) involves nearly 5 billion CSFs (4 940 906 560). However, in an appropriate MO basis, multiple excitations beyond quadruple ones often contribute very little to electron correlation of closed-shell species. Therefore adopting a RASSCF approach, the computational costs can be reduced to a reasonable level.³⁸

In our RASSCF calculations, the active space of M orbitals is divided into three subspaces, which are referred to as RAS1, RAS2, and RAS3.³⁵ They contain M_1 doubly occupied MOs, M_2 occupied or partially unoccupied MOs, and M_3 partially or unoccupied MOs, respectively. The total number of CSFs depends also on the chosen level of excitations (n_1, n_3) . Index n_1 means that up to n_1 electrons may be excited from the M_1 orbitals of RAS1 into RAS2 or RAS3 and up to n_3 electrons excited from RAS1 or RAS2 into the M_3 orbitals of RAS3. A suitably chosen combination of smaller n_1 and n_3 remarkably decreases the total number of CSFs. Then, a much larger active space of (N,M) can be used in RASSCF than in conventional CASSCF.

The RAS is here defined with reference to the ground-state MO energy level scheme of neutral UF_6 in Figure 1.^{3,22} The 18 highest occupied MOs with 36 electrons form RAS1, and the 7 lowest virtual MOs form RAS3, while RAS2 is left empty. Only single and double excitations were included, i.e., $n_1 = n_3 = 2$, since higher excitation levels improve the energy only at the 0.01 eV level. The obtained 1078 CSFs are a fraction of 0.22×10^{-6} of the CSF number of CAS(36,25). Among them, there are 126 excitations of one-electron singlet type and 126 of one-electron triplet type. In this centro-symmetric molecule, half of them are of spatial gerade (g) symmetry, the other half is ungerade (u). The orbital and state symmetry species are listed in Table 1.³⁹

Due to parity conservation, the RASSCF/RASPT2 calculations, without or with SOC on top of it, can be performed separately for all g- and u-states. First, state-average single-point RASSCF⁴⁰ calculations were carried out to obtain the wave functions for sets of 64 g-singlets (including the ground state) and 63 u-triplets. Then, RASPT2 calculations were performed for each state to improve the dynamic correlation energy. The intruder states problem caused converge troubles for several cases, which were resolved by a common energy-level shift, here by 0.3 au.^{41,42}

Table 2. Dominant AO Compositions (in %) and Energies (ϵ in eV) of the 18 Highest Occupied and 13 Lowest Virtual MOs of the UF_6 Ground State at the DFT-PW91 Scalar and SO Relativistic Levels

scalar relativistic (SR)			SO coupling	
orbital ^a	composition	ϵ	orbital ^b	ϵ
Highest Occupied				
1e _g	85 F-2p; 14 U-6d	-12.98	1u _{3/2g}	-12.98
1t _{2g}	88 F-2p; 12 U-6d	-12.33	2u _{3/2g}	-12.36
			1e _{5/2g}	-12.29
1t _{1u}	79 F-2p; 19 U-5f	-12.02	1e _{1/2u}	-12.04
			1u _{3/2u}	-12.01
1a _{1g}	97 F-2p; 1 U-7s	-11.75	1e _{1/2g}	-11.76
1t _{2u}	87 F-2p; 13 U-5f	-11.56	2u _{3/2u}	-11.56
			1e _{5/2u}	-11.55
1t _{1g}	100 F-2p	-10.72	2e _{1/2g}	-10.74
			3u _{3/2g}	-10.71
2t _{1u}	88 F-2p; 7 U-6p; 5 U-5f	-10.22	2e _{1/2u}	-10.73
			3u _{3/2u}	-9.93
Lowest Unoccupied				
1a _{2u}	100 U-5f	-7.08	2e _{5/2u}	-7.18
2t _{2u}	88 U-5f; 12 F-2p	-6.40	4u _{3/2u}	-6.36
			3e _{5/2u}	-6.26
3t _{1u}	88 U-5f; 12 F-2p	-5.11	5u _{3/2u}	-5.01
			3e _{1/2u}	-4.90
2a _{1g}	100 U-7s;	-0.77	3e _{1/2g}	-0.78
2t _{2g}	74 U-6d; 15 F-2p	-0.39	4u _{3/2g}	-0.56
			2e _{5/2g}	-0.11
2e _g	99 U-6d	+1.29	5u _{3/2g}	+1.29

^aLabeled by irreps of the O_h group. ^bLabeled by irreps the O_h double group.

Spin-orbit coupling was treated with the state interaction RAS/SO approach.⁴³ The SO matrix elements between the scalar relativistic RAS CSFs were evaluated with the SOC operator of the RECP formalism. For the diagonal elements of the SO interaction matrix, the RASPT2 state energies were used. In ample cases⁴⁴ including uranyl,³⁴ such an approach provides an accuracy comparable to all-electron-based SOC calculations.⁴⁵

The size of the SO matrix, coupling spin-singlet and spin-triplet CSFs, rises to 253×253 for the g-states, which becomes rather demanding. As a compromise we only included the lowest-lying 45 excited g-singlets and 45 g-triplets, together with the ground state, in the SO matrix for the g-states. So the dimension of the matrix was reduced to 181×181 . Similarly for the u-states, the lowest 48 u-singlets and 48 u-triplets were chosen, resulting in a SO matrix of 192×192 . Test calculations have shown that such a reduced RAS/SO scheme still gives a satisfactory accuracy.

All-electron relativistic TDDFT calculations, with or without SOC, were performed with the relativistic zeroth-order regular approximation (ZORA),^{46,47} as implemented in the ADF 2008 program.^{48–50} In the two-component scheme ZORA contains SOC terms of $(\sigma \cdot p V \sigma \cdot p)$ type. We used STO basis sets of triple- ζ plus polarization (TZP) quality for U and of DZP quality for F, from the ADF basis sets library. The SAOP^{51,52} and PW91⁵³ XC functionals were applied to compare the asymptote-corrected functional with common generalized gradient

approximation (GGA) functional. For comparison, a parallel series of TDDFT calculations with RECP was carried out by employing the NWChem 5.1 program^{54,55} with the same GTO basis sets and Stuttgart RECP as in the RAS calculations. In the NWChem calculations, we applied two pure density functionals (PW91 and BLYP)^{56,57} and two hybrid ones (B3LYP⁵⁸ and BHLYP)⁵⁹ to examine the effects of different XC functionals.

RESULTS AND DISCUSSION

Electronic States. The ground state of UF_6 is known to be $^1A_{1g}$ with valence electron configuration $\dots(1e_g)^4(1t_{2g})^6(1t_{1u})^6(1a_{1g})^2(1t_{2u})^6(1t_{1g})^6(2t_{1u})^6$ (Figure 1).²¹ The DFT PW91 energies and AO compositions of the occupied and virtual MOs at the scalar and SO relativistic levels are listed in Table 2. The 18 highest occupied MOs around -13 to -10 eV are close in energy and consist of the 2p lone pairs of the F^- ligands with up to 20% of U-5f or U-6d and less U-6p or U-7s. The 7 lowest unoccupied MOs ($1a_{2u}$, $2t_{2u}$, and $3t_{1u}$) around -7 to -5 eV are dominantly U-5f with a highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) gap of 2.75 eV with SOC, which is small for a closed-shell molecule. The high-lying virtual MOs ($2a_{1g}$, $2t_{2g}$, and $2e_g$) of dominant U-7s and U-6d character are more than 4 eV above the U-5f ones because they are pushed up more than the contracted U-5f ones by the filled F-2p shells owing to their stronger overlap (Figure 1). The low-energy occupied MOs, more than 10 eV below the active occupied ones, are already of semicore U-6p and F-2s type. Thus, only single electron excitations of LMCT type from the F^- 2p ligand shell to the U-5f shell occur in the region below 10 eV. The above-described RAS (2, 18, 2, 7) of 36 electrons and 18 + 7 orbitals with single and double excitations was thus chosen.

Spin-orbit coupling splits the triply degenerate t-type orbital levels of the ordinary O_h group into doubly and quadruply degenerate spinor levels of the double group: $t \otimes e_{1/2} = e_{1/2} \text{ or } 5/2 \oplus u_{3/2}$. DFT calculations show that, as expected, the SO splitting of the occupied MOs of F-2p type remains small, in the range of 0.1 eV. The only exception is the $2t_{1u}$ HOMO with a SO splitting of ~ 0.8 eV, due to the admixture²² of 7% U-6p (Table 2) with a large atomic SOC effect of U-6p in the order of 11 eV. The $2t_{1u}$ excited states exhibit even larger SO splittings (see below) because the U-6p admixture of $2t_{1u}$ increases by about one-half upon one-electron depletion of the $2t_{1u}$ orbital. Thereby, SOC reduces the HOMO–LUMO gap, leading to comparatively small excitation energies for the lowest states of g-type.²² Finally we have checked the flexibility of the basis sets by uncontracting the U-p functions,⁶⁰ and the energy changes of the SOC state energies were below 0.01 eV.

The double-group symmetries of the excited spin-orbit states are obtained in the usual manner. For instance, the first spin-averaged transitions ($2t_{1u} \rightarrow 1a_{2u}$) $^1A_{1g} \rightarrow ^3,^1T_{2g}$ generate a total of $(3 + 1) \times 3 = 12$ individual states, with two excitation energies of singlet–triplet and singlet–singlet types. Under SOC in the O_h^* double-group symmetry of a vibrationally undistorted UF_6 molecule, this splits up as follows (for more details, see Hay):²¹

$$u_{3/2u} \rightarrow e_{5/2u} : u_{3/2u} \otimes e_{5/2u} = E_g \oplus T_{1g} \oplus T_{2g}$$

$$e_{1/2u} \rightarrow e_{5/2u} : e_{1/2u} \otimes e_{5/2u} = A_{2g} \oplus T_{2g}$$

The calculated excitation energies of the g- and u-states at the SR- and SOC- RASPT2 levels up to nearly 10 eV are listed in the Supporting Information (Tables S1–S3). The excitation energies

Table 3. Theoretically Predicted Transition Energies (in eV) of Low-Lying Vertical Excited g-States of UF₆ in Comparison to Experimental Data

Scalar Relativistic Calculation				SOC Relativistic Calculation				Spectroscopic	
Orbital transition	El. state	This work ^a	Hay ^b	Spinor transition	El. state	Hay ^b	This work ^a	Solid UV Lewis ^c	El. Impact Cartwright ^d
$2t_{1u} \rightarrow 1a_{2u}$	$^3T_{2g}$	4.02	4.10	$(2t_{1u})3u_{3/2u} \rightarrow (1a_{2u})2e_{5/2u}$	$1T_{1g}$	3.35	3.33	3.046	} 3.22±0.01
	$^1T_{2g}$	4.27	4.55		$1E_g$	3.41	3.38		
					$1T_{2g}$	3.55	3.46	3.133	
$2t_{1u} \rightarrow 2t_{2u}$	$^3T_{2g}$	4.51	4.84	$(2t_{1u})3u_{3/2u} \rightarrow (2t_{2u})4u_{3/2u}$	$2T_{2g}$	4.14	4.04	3.761	} 4.14±0.02
	3E_g	4.54	4.54		$2E_g$	4.12	4.10		
	$^1T_{1g}$	4.71	4.91		$3T_{2g}$	4.39	4.14	} 3.848	
	$^3T_{1g}$	4.78	4.53		$2T_{1g}$	4.01	4.17		
	$^1T_{2g}$	4.79	4.99		$1A_{2g}$	4.36	4.25		
	$^1A_{2g}$	4.87	5.12		$3T_{1g}$	4.17	4.28	3.982	
	$^3A_{2g}$	4.97	5.10		$1A_{1g}$	4.35	4.30		
1E_g	5.00	5.25							
$2t_{1u} \rightarrow 3t_{1u}$	3E_g	5.50	5.41	$(2t_{1u})3u_{3/2u} \rightarrow (2t_{2u})3e_{5/2u}$	$4T_{2g}$	4.53	4.44	} 4.069	
	$^3A_{1g}$	5.57	5.52		$4T_{1g}$	4.43	4.47		
	$^3T_{1g}$	5.72	6.00		$3E_g$	4.56	4.53		
	$^1A_{1g}$	5.74	6.13	$(2t_{1u})2e_{1/2u} \rightarrow (1a_{2u})2e_{5/2u}$	$2A_{2g}$	5.83	5.03		
	$^1T_{1g}$	5.88	6.13		$5T_{2g}$	5.97	5.10		
	1E_g	5.88	6.61						
	$^3T_{2g}$	5.92	6.08						
	$^1T_{2g}$	5.92	6.21	$5T_{1g}$	5.24	5.28			
				$6T_{2g}$	5.24	5.33			
				$6T_{1g}$	5.28	5.34			
				$7T_{2g}$	5.59	5.38			
				$4E_g$	5.70	5.44			
				$2A_{1g}$	5.94	5.46			
				$3A_{2g}$	5.55	5.50			
Etc.									

^aRASPT2, this work. ^bHF + 1, Hay.²¹ ^cUV, solid phase, Lewis et al.¹⁵ ^dElectron impact, gas phase, Cartwright et al.¹⁶

of all g-states up to about 5 eV and of the T_{1u} states up to about 10 eV are here displayed in Tables 3 and 4, respectively, together with Hay's (HF + 1) results²¹ and the experimental data. Concerning the low-lying g-states, Hay's and our values (SR and SOC) agree within about 0.2 eV. However, for excited states above 5 eV, the discrepancy reaches 0.9 (0.7) eV for the SOC (SR) states of $2t_{1u} \rightarrow 3t_{1u}$ type, probably due to the limitation of electron correlation in Hay's CIS-like scheme. Accordingly, the rather dense energy-level order sometimes differs already for the low-lying states. Concerning the T_{1u} states, the agreement for the two lowest SR terms of $1t_{1g} \rightarrow 2t_{2u}$ and $1t_{1g} \rightarrow 3t_{1u}$ origin again is good (about 0.1 eV), while the discrepancies become larger for terms above 7 eV, where they reach 1.25 eV for $1e_g \rightarrow 3t_{1u}$. We attribute these enlarged discrepancies to the less reliable large-core pseudopotential used previously.²¹

Electronic Spectra. Experimental UV absorption spectra in condensed phase and electron-impact spectra in gas phase are known for UF₆ up to about 10 eV.^{15,16,61} They involve mostly

electric dipole allowed $A_{1g} \rightarrow T_{1u}$ transitions. Weaker transitions observable at energies below the onset of the strong ones may be magnetic dipole allowed ($A_{1g} \rightarrow T_{1g}$) or vibronically induced, in particular by the t_{1u} and t_{2u} bending vibrations, relevant for F-t_{1u} → U-Sf excitations. Below 4.5 eV, 6 sharp UV absorption lines were observed in the solid (Table 3) and can be tentatively assigned to final T_{1,2g} states, which occur 0.3 eV lower than calculated by RAS/SO. In the same energy range, two broader molecular electron-impact lines were observed. The corresponding groups of calculated transition energies lie at most about 0.1 eV higher. We note the large red shift of the solid-state spectra of about 0.2 eV, which is reminiscent of similarly large noble gas matrix effects for CUO and UO₂ molecules.^{62–65}

The lowest g → g excitations near 3 eV originate from $(2t_{1u})3u_{3/2u} \rightarrow (1a_{2u})2e_{5/2u}$ orbital transitions. Compared to the SR energies, they are strongly spin-orbit stabilized by about 0.7 eV. The $(2t_{1u})5e_{1/2u} \rightarrow (1a_{2u})2e_{5/2u}$ spin-orbit mates are calculated

Table 4. Theoretically Predicted Excitation Energies (ΔE , in eV) and Transition Moments (μ , in Debye) of Low-Lying T_{1u} States of UF_6 in Comparison to Experimental Transition Energies

Orbital transition	Scalar		Spin-Orbit Coupled		Spectroscopic El.Impact Cartwright ^c	
	This work ^a	Hay ^b	Hay ^b	This work ^a		
$1t_{1g} \rightarrow 1a_{2u}$	-	-	5.39 (0.43)	4.98 (0.47)	4.77±0.18	
			6.01 (0.38)	5.63 (0.45)	5.41±0.06	
$1t_{1g} \rightarrow 2t_{2u}$	5.95 (1.18)	6.10 (1.58)	6.15	5.72 (0.13)	5.87±0.03	
			6.22 (0.69)	5.76 (0.98)		
			6.40 (1.04)	6.00 (0.92)		
$1a_{1g} \rightarrow 2t_{2u}$	-	-	6.85 (0.67)	6.26 (0.70)		
					7.05	6.82 (0.46)
$1t_{1g} \rightarrow 3t_{1u}$	6.87 (0.26)	6.82 (0.43)	7.12 (0.51)	6.84 (0.33)	7.09±0.08	
			7.25 (0.46)	6.98 (0.79)		
			7.63	7.17 (0.24)		
			7.65	7.19 (0.28)		
$1t_{2g} \rightarrow 1a_{2u}$	7.22 (0.05)	7.40 (0.30)	7.35 (0.30)	6.87 (0.45)		
			7.41 (0.23)	6.92 (0.08)		
$1a_{1g} \rightarrow 3t_{1u}$	7.60 (4.15)	8.35 (3.08)	7.90 (0.33)	7.50 (0.97)	7.92±0.04	
			8.75 (3.02)	7.86 (1.56)		
$1t_{2g} \rightarrow 2t_{2u}$	7.74 (0.03)	7.97 (0.79)	8.06	7.67 (0.97)		
			8.10	7.73 (0.33)		
			8.23	7.84 (0.35)		
			8.36	7.88 (0.13)		
$1e_g \rightarrow 1a_{2u}$	-	-	8.38 (0.28)	7.90 (0.01)		
					7.96	7.07 (0.35)
$1e_g \rightarrow 2t_{2u}$	7.81 (0.01)	8.63 (1.12)	8.48 (0.61)	7.64 (1.70)		
			8.70 (0.74)	7.97 (0.74)		
$1t_{2g} \rightarrow 3t_{1u}$	8.75 (0.35)	9.10 (0.58)	8.99 (0.76)	8.06 (0.24)		
					9.31	
					9.38 (0.51)	
$1e_g \rightarrow 3t_{1u}$	9.02 (7.66)	10.27 (4.65)	9.41			
					9.49 (0.48)	
					10.20	
			10.38 (0.48)			
			10.65 (4.45)		9.13±0.04	

^a RASPT2, this work. States above 8.5 eV were excluded from the SOC calculations. ^b HF + 1, Hay.²¹ ^c Electron impact, experimental.¹⁶

1.7 eV higher and are no longer observable under the onset of the strongly allowed transitions. The SO splitting of 0.09 eV (701 cm^{-1}), assigned to $2t_{1u}$ in the UV matrix work,¹⁵ seems too small by more than an order of magnitude. We interpret this small splitting as due to differential F-2p/U-5f Coulomb coupling restrained by U-6p SOC. The enlarged F-2p ($2t_{1u}$) SO splitting has already been discussed above and is also consistent with the photoelectron spectra (PES) of gaseous UF_6 (Figures 1 and 2 of ref 66). From there, a ($2t_{1u}$) $u_{3/2u} - e_{1/2u}$ SO-splitting of about 1.4–1.5 eV can be deduced. The second group of weak features near 4 eV emerges from ($2t_{1u}$) $3u_{3/2u} \rightarrow (2t_{2u}) 4u_{3/2u} 3e_{5/2u}$. We attribute the comparatively large U-5f (t_{2u}) final-orbital SO-splitting of about 0.2 eV to the large effective positive charge of U in UF_6 .

The calculated electric dipole-allowed T_{1u} excitation energies and transition moments are listed in Table 4 and are compared to

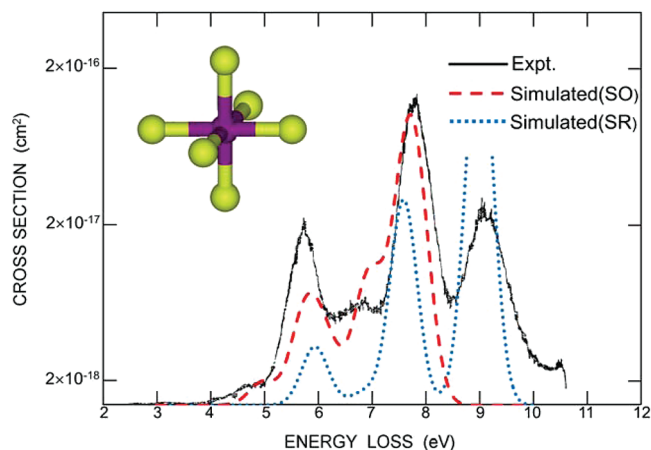


Figure 2. Experimental and theoretically simulated electric dipole-allowed electronic spectrum of UF_6 . Solid curve: electron impact forward scattering,¹⁶ dotted curve: SR calculation up to 9.5 eV; dashed curve: spin-orbit (SO) coupled calculation up to 8.5 eV.

the high-energy (75 eV) forward-scattering (5°) electron-impact features.¹⁶ Our RASPT2/SO data reproduce the peak energies within about 0.2 eV. The lowest two T_{1u} states are calculated at 4.98 and 5.63 eV (corresponding to 5.39 and 6.01 eV from Hay's work)²¹ with transition moments below 0.5 D and are hereby assigned as weak electric dipole-allowed spectral features. Concerning the experimental feature at 4.9 eV, our results show that it is not a forbidden transition, supporting the experimental assignment. The SOC wave function of the lowest T_{1u} state, for instance, is dominated by ($1t_{1g} \rightarrow 1a_{2u}$) $^3T_{2u}$; without SOC the transition would be singlet-triplet forbidden and 0.3 eV higher.

Figure 2 shows the comparison of the experimental electron-impact forward-scattering spectrum¹⁶ and our SR- and SO-RASPT2 simulations. The 24 calculated T_{1u} transitions are represented by Gaussian functions with heights proportional to the squared transition moments and a full width at the half-maximum (fwhm) of 0.3 eV. The positions and relative intensities of the experimental features are reasonably reproduced by the SOC simulation. This gives credence to the validity of the slightly revised spectral assignments in Table 4. The simulation also explains why only 9 maxima were experimentally resolvable from the overlay of dozens of $g \rightarrow g$ and $g \rightarrow u$ transitions. While SOC shifts the peak positions of the SR curve only a little, it markedly changes the intensity pattern. The SR curve reproduces the position of the strong band near 9 eV but misses the weak peaks and shoulders below 5.5 eV. Our recent work also shows that applying this SO-RASPT2 approach, one can accurately predict the fluorescent states of UO_2F_2 and $UO_2F_2(Ar)_x$ complexes.⁶⁷

Dependence of RASPT2 Accuracy upon the RAS Level. As the accuracy of RASPT2 approaches depend on the type and number of allowed excitations in the RASSCF step, determined by the four parameters (n_1, M_1, n_3, M_3), we investigated various excitation levels (n_1, n_3). In as much as the (36,25) space is too large for performing comparable RASCF and CASSCF calculations, we carried out the test calculations with a (14,14) space. In such a size-reduced scheme, the number of CSFs in the CASSCF is 348 615, which is only 1/14000 of the CSFs number in CAS(36,25). For the 14 active orbitals, we selected the 7 upper occupied orbitals ($1a_{1g}$, $1t_{1g}$, and $2t_{1u}$) of nonbonding F-2p-type and the 7 low-lying empty orbitals ($1a_{2u}$, $2t_{2u}$, and $3t_{1u}$) of U-5f

Table 5. SR RASPT2 Excitation Energies (in meV) of UF₆

state ^a	Δ-RAS at level <i>n</i> ^b						14 = CAS
	<i>n</i> = 1	2	3	4	5	7	
1 ¹ T _{2u}	-295.9	+38.0	+7.3	-0.2	+0.1	+0.0	4681.1
2 ¹ T _{2u}	-251.1	+46.1	+8.8	-0.1	+0.2	+0.0	5304.7
1 ¹ T _{1u}	-178.7	+41.6	+6.8	-0.2	+0.1	+0.0	5356.3

^a For the three lowest electric dipole-allowed transitions to 1¹T_{2u}, 2¹T_{2u}, and 1¹T_{1u}. ^b Deviation at RAS levels (*n*,7,*n*,7) from the CAS(14,14) values (*n* = 14).

Table 6. NOON from SR MCSCF Calculations of UF₆ and UO₂²⁺ Ground States Using the R²ASSCF Scheme

UF ₆ ^a	active NO	1e _g	1t _{2g}	1t _{1u}	1t _{2u}	2t _{2u}	3t _{1u}	2t _{2g}	2e _g
	occ. no.	1.988	1.991	1.992	1.981	0.020	0.006	0.009	0.012
UO ₂ ²⁺ ^b	active NO	π _g	σ _g	π _u	σ _u	π _u *	σ _u *	π _g *	σ _g *
	occ. no.	1.973	1.973	1.962	1.976	0.039	0.038	0.026	0.013

^a CAS(22,22)/RAS(2,11,2,11). ^b CAS(12,12)/RAS(2,6,2,6).

type. Choosing $M_1 = M_3 = 7$ and $M_2 = 0$ for the subspace partitioning, we varied $n_1 = n_3 = 1, 2, 3, \dots, 14$, where for $n = 14$, the RAS reaches the complete active space (CAS).

For tests at the SR level, we investigated the 1¹A_{1g} ground state and three low-lying excited u-states (1¹T_{2u}, 2¹T_{2u}, and 1¹T_{1u}). The dependence of the excitation energies on parameter n is displayed in Table 5. Remarkably, the RASPT2 excitation energies quickly converge toward the CASPT2 results: For $n = 2$ (i.e., up to double excitations), the maximal absolute error is less than 0.05 eV, and it drops below 0.2 meV for $n \geq 4$. Since the computational cost increases dramatically with the excitation index, while the accuracy gain drops drastically, restriction to single and double excitations seems a good compromise for the present cases of LMCT transitions from ligand lone pairs to metal nonbonding orbitals of an otherwise closed-shell complex. The RASSCF approach with single and double excitations (hereafter denoted by R²ASSCF) and additional RASPT2 calculations will be applied throughout this article, if not specified otherwise.

To obtain accurate results of the excitation energies, an important question is regarding how far occupied U-6p and virtual U-6d/7s orbitals are needed in the active space. Bonding and antibonding MOs involving U-6d and U-5f contribute considerably to correlation in the uranyl cation (UO₂²⁺) and are usually included in the active space.^{12,68} We have attempted to compare the U-6d contributions to the RASSCF results of UF₆ and UO₂²⁺ in their ground states with the R²ASSCF scheme. For UF₆, an active space of (22,22) was constructed from the 11 occupied valence MOs that contain U-6d (1e_g, 1t_{2g}) and U-5f (1t_{1u}, 1t_{2u}) contributions, and their 11 unoccupied counterparts (2t_{2u}, 3t_{1u} and 2t_{2g}, 2e_g). The (12,12) active orbital set for UO₂²⁺ is formed by the 6 pairs of bonding/antibonding MOs, i.e., σ_u/σ_u*, σ_g/σ_g*, π_u/π_u*, and π_g/π_g*. The natural orbital occupation numbers (NOON) of the two molecular species are listed in Table 6.

As an approximate rule of thumb, for accurate multireference results natural orbitals with occupation number larger than 0.01 should be included in the CAS or RAS.^{11,69,70} Table 6 shows that all 6 or at least 5 antibonding MOs of the triply bonded [O≡U≡O]²⁺ should be included in the CAS as their NOONs

Table 7. SR Excitation Energies (in eV) for Low Singlets of UF₆ from SR-TDDFT Using Different Basis Sets and Various XC Functionals Compared to RASPT2 Results

excited state	TDDFT						RASPT2 with RECP ^d
	RECP ^a				all-electron ZORA ^b		
	BLYP	PW91	B3LYP	BHLYP	PW91	SAOP	
g States							
1 ¹ T _{2g}	3.16	3.17	3.83	4.81	3.27	3.88	4.27
1 ¹ T _{1g}	3.73	3.73	4.21	5.00	3.82	4.54	4.71
2 ¹ T _{2g}	3.79	3.80	4.25	5.02	3.91	4.62	4.79
1 ¹ A _{2g}	3.74	3.74	4.26	5.19	3.82	4.54	4.87
1 ¹ E _g	4.07	4.08	4.57	5.43	4.21	4.88	5.00
2 ¹ T _{1g}	4.25	4.28	5.06	6.54	4.47	4.98	5.49
2 ¹ A _{1g}	5.00	5.02	5.37	6.13	5.23	5.84	5.74
3 ¹ T _{1g}	4.94	4.97	5.35	6.06	5.16	5.76	5.88
2 ¹ E _g	4.98	5.01	5.57	6.35	5.22	5.81	5.88
3 ¹ T _{2g}	4.93	4.95	5.45	6.11	5.15	5.77	5.92
ave dev ^b	-1.00	-0.98	-0.46	+0.41	-0.83	-0.21	
max dev ^c	-1.24	-1.21	-0.61	+1.05	-1.05	-0.51	
u States							
1 ¹ T _{2u}	3.45	3.44	4.38	5.97	3.63	4.25	5.24
1 ¹ A _{2u}	4.19	4.18	4.93	6.33	4.40	5.09	5.69
1 ¹ E _u	4.14	4.13	4.89	6.29	4.33	5.05	5.70
2 ¹ A _{2u}	4.53	4.51	5.32	6.67	4.70	5.24	5.72
2 ¹ T _{2u}	4.16	4.15	4.95	6.41	4.35	5.06	5.81
1 ¹ T _{1u}	4.37	4.36	5.16	6.63	4.57	5.25	5.97
2 ¹ T _{2u}	5.08	5.06	5.86	7.02	5.40	6.04	6.30
3 ¹ T _{2u}	5.22	5.20	5.95	7.03	5.69	6.30	6.81
2 ¹ T _{1u}	5.39	5.36	5.99	7.07	5.74	6.35	6.87
ave dev ^c	-1.51	-1.52	-0.74	+0.55	-1.24	-0.61	
max dev ^d	-1.79	-1.80	-0.88	+0.95	-1.61	-0.75	

^a Same small-core potentials and basis sets as for RASPT2. ^b STO basis sets. ^c Average deviation from the RASPT2 results. ^d Maximal deviation from the RASPT2 results.

exceed 0.025. The demand on the effective active space is less serious in the more ionically bonded U(-F)₆. With NO occupations around 0.01, the omission of U-6d type 2t_{2g} and 2e_g MOs from the active space should not lead to significant errors. Also the exclusion of U-6p orbitals from the RAS seems an acceptable compromise.

Evaluation of TDDFT. Even with simplifications by RASPT2 schemes, multiconfiguration WFT calculations are still expensive for actinide compounds, such as UF₆. It is thus worth investigating the quality of DFT and TDDFT approaches in the evaluation of the excited-state properties. Therefore, we have performed TDDFT calculations at the SR level using the all-electron ZORA and the valence-electron RECP approaches. The excitation energies for a number of low-lying excited singlets using different XC-functionals are shown in Table 7.

Obviously, the calculated excited states energies of UF₆ with TDDFT have significant errors when comparing to RASPT2 and experiments, at least concerning (F-2p)₆ → U-5f type excitations involving LMCT from extended orbitals to compact orbitals. All-electron ZORA and valence-electrons RECP results exhibit similar error patterns. For the PW91 functional, the two schemes do not differ by more than 0.2 eV. The two GGA functionals BLYP

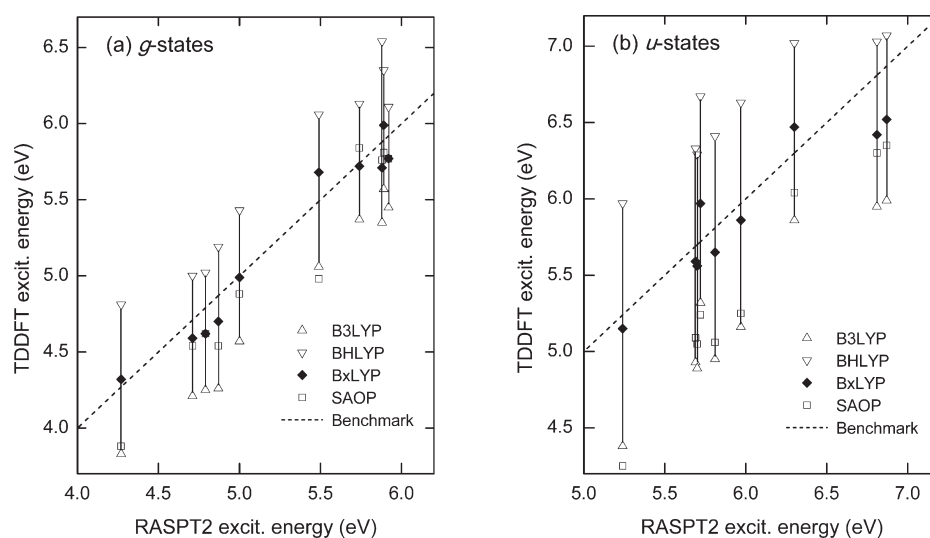


Figure 3. SR TDDFT excitation energies obtained by SAOP, BxLYP (with $\alpha = 35\%$ HF exchange), B3LYP, and BHLYP.

and PW91 produce similar deviations, around -1 eV for the g-states and around -1.5 eV for the u-states. The two hybrid functionals B3LYP and BHLYP reduce the discrepancies to about one-half. Even the asymptotically corrected SOAP functional generates errors for the ungerade $F-2p \rightarrow U-5f$ type excitation energies of about -0.6 eV. Wahlgren et al. had speculated that the errors of TDDFT for the excited states may be due to the contributions from double excitations,³³ which are not accounted for by TDDFT. Among other error sources there is the incorrect asymptotic behavior of many XC functionals,⁷¹ which leads to a typical underestimation of most excitation energies. This can be alleviated by XC functionals with adjusted asymptotic behavior like the SAOP. However, the SAOP functional still does not provide a satisfactory description of the LMCT states of UF_6 .

The TDDFT energies of these charge-transfer transitions to very compact U-5f MOs seem particularly sensitive to the percentage of HF exchange in hybrid XC-potentials.⁷² We note that adjusting the admixture of the HF exchange alleviates the DFT problems in the present case. As shown in Table 7, the HF exchange percentage increase from 20% (in B3LYP) to 50% (in BHLYP) turns the average deviation from -0.46 into $+0.41$ for the g-states and from -0.74 into $+0.55$ eV for the u-states. We therefore form a hybrid functional BxLYP with intermediate $\alpha = 35\%$ HF exchange, which reproduces the RASPT2 excitation energies of UF_6 quite well (Figure 3). Further work is needed to investigate if this simple approach can help to predict the excited-states energies of other actinide compounds.

SUMMARY AND CONCLUSIONS

The electronic excited states involving transitions from F lone pairs to U-5f in UF_6 have been investigated by using an ab initio multireference electron correlation method. With the RASSCF-based approach, a single and double excitation restriction is shown to be acceptable for this type of transitions. With the Stuttgart small-core relativistic ECP and adapted one-electron basis sets, the spin-orbit interaction important for some of the transitions due to U-6p semicore admixture is accounted for by a RECP-SOC pseudopotential operator and the RASSCF/RASPT2/SO technique. Dynamic electron correlations are well recovered by second-order perturbation theory RASPT2. This

scheme reproduces the spectral energies (within about 0.2 eV) and the intensities of the broad experimental electron-impact peaks below 10 eV. These transitions are of mixed singlet-triplet type. The two lowest-energy features are only vibronically and/or magnetic-dipole allowed, the next seven features are optically dipole allowed. Assuming a condensed phase red shift of 0.2 eV, the energies of the UV-absorption peaks in the optically forbidden region below 4.5 eV can also be well predicted. Given the numerous ligand lone pairs involved in the low-lying excited states, the quantitative prediction of the electronic spectra of UF_6 is computationally intensive. Our results represent so far the best theoretical reproduction of the few measured and the prediction of a large number of not yet detected excited states of UF_6 . An improved tentative assignment of the measured excitations is given in Tables 3 and 4.

The presented RASSCF/RASPT2/SO technique using RECP appears as a promising tool for the investigation of the excited states of actinide compounds. It can serve as a benchmark for calibrating approximation approaches, such as SO-TDDFT, for applications in heavy-element systems. Even the asymptotically corrected SAOP functional is not yet sufficiently accurate for these comparatively simple one-electron excitations of actinide complexes. New exchange-correlation functionals that can handle charge-transfer and compact actinide 5f-states are needed for application of DFT and TDDFT for predicting the excited-states properties of actinide systems.

ASSOCIATED CONTENT

S Supporting Information. Extended lists of excited-state energies are available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: junli@mail.tsinghua.edu.cn.

Present Addresses

[†]Permanent address: Physical and Theoretical Chemistry, Faculty of Science and Engineering, Siegen University, Siegen 57068, Germany. E-mail: schwarz@chemie.uni-siegen.de.

ACKNOWLEDGMENT

W.H.E.S. thanks for the hospitality of the Theoretical and Computational Chemistry Laboratory (TCCL). We acknowledge the financial support from NKBRFS (2011CB932400) and NSFC (20933003, 11079006, 91026003) of China. The calculations were performed using computers at the Supercomputing Center of the Computer Network Information Center, Chinese Academy of Sciences, Tsinghua National Laboratory for Information Science and Technology, and the Shanghai Supercomputing Center.

REFERENCES

- (1) Hay, P. J.; Martin, R. L. *J. Chem. Phys.* **1998**, *109*, 3875–3881.
- (2) Gagliardi, L.; Willetts, A.; Sklyaris, C.-K.; Handy, N. C.; Spencer, S.; Ioannou, A. G.; Simper, A. M. *J. Am. Chem. Soc.* **1998**, *120*, 11727–11731.
- (3) Xiao, H.; Li, J. *Chin. J. Struct. Chem.* **2008**, *27*, 967–974.
- (4) (a) Hu, S.-W.; Wang, X.-Y.; Chu, W.-W.; Liu, X.-Q. *J. Phys. Chem. A* **2008**, *112*, 8877–8883. (b) Hu, S.-W.; Wang, X.-Y.; Chu, W.-W.; Liu, X.-Q. *J. Phys. Chem. A* **2009**, *113*, 9243–9248.
- (5) Garrison, S. L.; Becnel, J. M. *J. Phys. Chem. A* **2008**, *112*, 5453–5457.
- (6) De Jong, W. A.; Nieuwpoort, W. C. *Int. J. Quantum Chem.* **1996**, *58*, 203–216.
- (7) Garcia-Hernández, M.; Lauterbach, C.; Krüger, S.; Matveev, A.; Röscher, N. *J. Comput. Chem.* **2002**, *23*, 834–846.
- (8) Larsson, S.; Tse, J. S. *Chem. Phys.* **1984**, *89*, 43–50.
- (9) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483–5488.
- (10) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (11) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094.
- (12) Pierloot, K.; Besien, E. *J. Chem. Phys.* **2005**, *123*, 204309.
- (13) Küchle, W.; Dolg, M.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1994**, *100*, 7535–7542.
- (14) Vallet, V.; Macak, P.; Wahlgren, U.; Grenthe, I. *Theor. Chem. Acc.* **2006**, *115*, 145–160.
- (15) Lewis, W. B.; Asprey, L. B.; Jones, L. H.; McDowell, R. S.; Rabideau, S. W.; Zeltmann, A. H.; Paine, R. T. *J. Chem. Phys.* **1976**, *65*, 2707–2714.
- (16) Cartwright, D. C.; Trajmar, S.; Chutjian, A.; Srivastava, S. *J. Chem. Phys.* **1983**, *79*, 5483–5493.
- (17) Onoe, J.; Takeuchi, K.; Nakamatsu, H.; Mukoyama, T.; Sekine, R.; Adachi, H. *Chem. Phys. Lett.* **1992**, *196*, 636–640.
- (18) Malli, G. L.; Styszynski, J. *J. Chem. Phys.* **1996**, *104*, 1012–1017.
- (19) Hay, P. J.; Wadt, W. R.; Kahn, L. R.; Raffanetti, R. C.; Phillips, D. H. *J. Chem. Phys.* **1979**, *71*, 1767–1779.
- (20) Evestov, R. A.; Panin, A. I.; Bandura, A. V. *Russ. J. Gen. Chem.* **2008**, *78*, 1823–1835.
- (21) Hay, P. J. *J. Chem. Phys.* **1983**, *79*, 5469–5482.
- (22) Xiao, H.; Hu, H.-S.; Schwarz, W. H. E.; Li, J. *J. Phys. Chem. A* **2010**, *114*, 8837–8844.
- (23) Malmqvist, P.-Å.; Pierloot, K.; Moughal Shahi, A. R.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.
- (24) Moughal Shahi, A. R.; Cramer, C. J.; Gagliardi, L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10964–10972.
- (25) Sauri, V.; Serrano-Andrés, L.; Moughal Shahi, A. R.; Gagliardi, L.; Vancoillie, S.; Pierloot, K. *J. Chem. Theory Comput.* **2011**, *7*, 153–168.
- (26) Gagliardi, L.; Heaven, M. C.; Krogh, J. W.; Roos, B. O. *J. Am. Chem. Soc.* **2005**, *127*, 86–91.
- (27) Infante, I.; Andrews, L.; Wang, X.; Gagliardi, L. *Chem.—Eur. J.* **2010**, *16*, 12804–12807.
- (28) (a) Berning, A.; Schweizer, M.; Werner, H.-J.; Knowles, P. J.; Palmieri, P. *Mol. Phys.* **2000**, *98*, 1823. (b) Figgen, D.; Müller, W.; Schweizer, M.; Stoll, H.; Peterson, K. A. *J. Chem. Phys.* **2003**, *118*, 4766–4767. (c) Peterson, K. A.; Shepler, B. C.; Singleton, J. M. *Mol. Phys.* **2007**, *105*, 1139–1155.
- (29) Hafner, P.; Schwarz, W. H. E. *Chem. Phys. Lett.* **1979**, *65*, 537–541.
- (30) Han, Y. K.; Hirao, K. *J. Chem. Phys.* **2000**, *113*, 7345–7350.
- (31) Hay, P. J.; Martin, R. L. *J. Chem. Phys.* **1998**, *109*, 3875–3881.
- (32) Peralta, J. E.; Batista, E. R.; Scuseria, G. E. *J. Chem. Theory Comput.* **2005**, *1*, 612–616.
- (33) Réal, F.; Vallet, V.; Marian, C.; Wahlgren, U. *J. Chem. Phys.* **2007**, *127*, 214302.
- (34) Wei, F.; Wu, G.-S.; Schwarz, W. H. E.; Li, J. *Theor. Chem. Acc.* **2011**, *129*, 467–481.
- (35) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardtsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Koppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2008.1, a package of ab initio programs; University College Cardiff Consultants Limited: Wales, U.K., 2008; see <http://www.molpro.net>.
- (36) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (37) Seip, H. M. *Acta Chem. Scand.* **1965**, *20*, 2698–2710.
- (38) Olsen, J.; Roos, B. O.; Jørgensen, P.; Jensen, H. J. A. *J. Chem. Phys.* **1988**, *89*, 2185–2192.
- (39) Altmann, S. L.; Herzog, P. *Poincaré-Group Theory Tables*; Oxford University Press: Oxford, UK, 1994.
- (40) Werner, H. J.; Meyer, W. *J. Chem. Phys.* **1981**, *74*, 5794–5801.
- (41) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215–223.
- (42) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M.; Molina, V. *J. Mol. Struct. Theochem* **1998**, *388*, 257–276.
- (43) Malmqvist, P.-Å.; Roos, B. O.; Schimmelpfennig, B. *Chem. Phys. Lett.* **2002**, *357*, 230–240.
- (44) Schimmelpfennig, B.; Maron, L.; Wahlgren, U.; Teichteil, C.; Fagerli, H.; Gropen, O. *Chem. Phys. Lett.* **1998**, *286*, 267–271.
- (45) Ramírez-Solís, A.; Daudey, J. P. *J. Chem. Phys.* **2004**, *120*, 3221–3228.
- (46) Chang, C.; Pelissier, M.; Durand, M. *Phys. Scri.* **1986**, *34*, 394–404.
- (47) Faas, S.; Snijders, J. G.; van Lenthe, J. H.; van Lenthe, E.; Baerends, E. J. *Chem. Phys. Lett.* **1995**, *246*, 632–640.
- (48) te Velde, G.; Bickelhaupt, F. M.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (49) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.
- (50) *ADF2009.01*, SCM, Theoretical Chemistry; Vrije Universiteit: Amsterdam, 2009; <http://www.scm.com>
- (51) Gritsenko, O. V.; Schipper, P. R. T.; Baerends, E. J. *Chem. Phys. Lett.* **1999**, *302*, 199–207.
- (52) Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2000**, *112*, 1344–1352.
- (53) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- (54) Straatsma, T. P.; Aprà, E.; Windus, T. L.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Harrison, R.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtel, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dylla, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, version 5.1 (2006); Pacific Northwest National Laboratory: Richland, Washington, 2006.
- (55) Kendall, R. A.; Aprà, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.;

Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260–283.

- (56) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (57) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (58) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (59) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (60) Weigend, F.; Baldes, A. *J. Chem. Phys.* **2010**, *133*, 174102 (10pp).
- (61) (a) Bernstein, E. R.; Meredith, G. R. *Chem. Phys.* **1977**, *24*, 289–299, 301–309, 311–325. (b) Rice, W. W.; Oldenborg, R. C.; Wantuck, P. J.; Tjee, J. J.; Wampler, F. P. *J. Chem. Phys.* **1980**, *73*, 3560–3567. (c) Baranov, V. Y.; Kolesnikov, Y. A.; Kotov, A. A. *Quant. Electron.* **1999**, *29*, 653–666.
- (62) Andrews, L.; Liang, B.; Li, J.; Bursten, B. E. *Angew. Chem., Int. Ed.* **2000**, *39*, 4565–4567.
- (63) Li, J.; Bursten, B. E.; Liang, B.; Andrews, L. *Science* **2002**, *295*, 2242–2245.
- (64) Li, J.; Bursten, B. E.; Liang, B.; Andrews, L. *J. Am. Chem. Soc.* **2002**, *124*, 9016–9017.
- (65) Li, J.; Bursten, B. E.; Andrews, L.; Marsden, C. J. *J. Am. Chem. Soc.* **2004**, *126*, 3424–3425.
- (66) Mårtensson, N.; Malmquist, P. Å.; Svensson, S.; Johansson, B. *J. Chem. Phys.* **1984**, *80*, 5458–5464.
- (67) Su, J.; Wang, Y.-L.; Wei, F.; Schwarz, W. H. E.; Li, J. *J. Chem. Theory Comp.* **2011**; DOI: 10.1021/ct200419x.
- (68) Gagliardi, L.; Roos, B. O. *J. Phys. Chem. A* **2001**, *105*, 10602–10606.
- (69) (a) Pulay, P.; Hamilton, T. P. *J. Chem. Phys.* **1988**, *88*, 4926–4933. (b) Bofill, J. M.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3637–3646; (c) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3647–3659.
- (70) Gordon, M. S.; Schmidt, M. W.; Chaban, G. M.; Glaesemann, K. R. *J. Chem. Phys.* **1999**, *110*, 4199–4207.
- (71) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.
- (72) Autschbach, J. *ChemPhysChem* **2009**, *10*, 1757–1760.

Calibration of DFT Functionals for the Prediction of ^{57}Fe Mössbauer Spectral Parameters in Iron–Nitrosyl and Iron–Sulfur Complexes: Accurate Geometries Prove Essential

Gregory M. Sandala,[†] Kathrin H. Hopmann,[‡] Abhik Ghosh,[‡] and Louis Noodleman^{*,†}

[†]Department of Molecular Biology, TPC15, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037, United States

[‡]Center of Theoretical and Computational Chemistry and Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway

S Supporting Information

ABSTRACT: Six popular density functionals in conjunction with the conductor-like screening (COSMO) solvation model have been used to obtain linear Mössbauer isomer shift (IS) and quadrupole splitting (QS) parameters for a test set of 20 complexes (with 24 sites) comprised of nonheme nitrosyls (Fe–NO) and non-nitrosyl (Fe–S) complexes. For the first time in an IS analysis, the Fe electron density was calculated both directly at the nucleus, $\rho(0)_{\text{N}}$, which is the typical procedure, and on a small sphere surrounding the nucleus, $\rho(0)_{\text{S}}$, which is the new standard algorithm implemented in the ADF software package. We find that both methods yield (near) identical slopes from each linear regression analysis but are shifted with respect to $\rho(0)$ along the x axis. Therefore, the calculation of the Fe electron density with either method gives calibration fits with equal predictive value. Calibration parameters obtained from the complete test set for OLYP, OPBE, PW91, and BP86 yield correlation coefficients (r^2) of approximately 0.90, indicating that the calibration fit is of good quality. However, fits obtained from B3LYP and B3LYP* with both Slater-type and Gaussian-type orbitals are generally found to be of poorer quality. For several of the complexes examined in this study, we find that B3LYP and B3LYP* give geometries that possess significantly larger deviations from the experimental structures than OLYP, OPBE, PW91, or BP86. This phenomenon is particularly true for the di- and tetranuclear Fe complexes examined in this study. Previous Mössbauer calibration fit studies using these functionals have usually included mononuclear Fe complexes alone, where these discrepancies are less pronounced. An examination of spin expectation values reveals that B3LYP and B3LYP* approach the weak-coupling limit more closely than the GGA exchange-correlation functionals. The high degree of variability in our calculated S^2 values for the Fe–NO complexes highlights their challenging electronic structure. Significant improvements to the isomer shift calibrations are obtained for B3LYP and B3LYP* when geometries obtained with the OLYP functional are used. In addition, greatly improved performance of these functionals is found if the complete test set is grouped separately into Fe–NO and Fe–S complexes. Calibration fits including only Fe–NO complexes are found to be excellent, while those containing the non-nitrosyl Fe–S complexes alone are found to demonstrate less accurate correlations. Similar trends are also found with OLYP, OPBE, PW91, and BP86. Correlations between experimental and calculated QSs were also investigated. Generally, universal and separate Fe–NO and Fe–S fit parameters obtained to determine QSs are found to be of good to excellent quality for every density functional examined, especially if $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$ is removed from the test set.

INTRODUCTION

Mössbauer (MB) spectroscopy is a valuable experimental tool in bioinorganic chemistry because it can provide information on the spin and oxidation states of the MB atom and fine structure parameters like zero-field splitting and the hyperfine coupling tensor.¹ By far, ^{57}Fe represents the most commonly probed nucleus, though it is worth noting that ^{61}Ni and ^{67}Zn are two other biologically relevant MB-active isotopes. MB spectroscopy relies on the recoilless resonant absorption of γ radiation by the MB active nuclei from a source emitter. For ^{57}Fe , it is the resonant absorption of 14.4 keV γ -rays emitted from a radioactive ^{57}Co source that is measured. An advantage of MB spectroscopy is that all MB active nuclei are probed simultaneously in the sample. Challenges in using this technique include the requirement of ^{57}Fe (or other) enrichment and deconvoluting complicated spectra when several active nuclei are present. In conjunction with rapid freeze-quench techniques, MB spectroscopy is a particularly

valuable tool in bioinorganic chemistry because the Fe environment can be monitored during the course of a biochemical reaction to give insight into the nature of short-lived intermediates and the reaction mechanisms in which they participate.²

Two key parameters extracted from Mössbauer spectra are the isomer shift (IS) and quadrupole splitting (QS). ISs provide information on metal–ligand covalency and the spin and oxidation states of the Fe center. Contributions to the magnitude of ISs are effects from shielding of the 3s electrons by the Fe 3d electrons, covalency effects, and changes in bond lengths.³ The IS can be calculated by linear regression of the following equation:

$$\delta = \alpha[\rho(0) - A] + C \quad (1)$$

Received: March 19, 2011

Published: August 31, 2011

where δ is the experimental IS (mm s^{-1}), $\rho(0)$ is the calculated s-electron density at the nucleus ($e a_0^{-3}$), α is the slope ($e^{-1} a_0^3 \text{mm s}^{-1}$), and C is the intercept (mm s^{-1}). A is a large constant chosen to be near $\rho(0)$ in magnitude. This aids in the numerical stability of the fitting in the linear regression equation. The factors α , A , and C are determined in calibration procedures, often with calculations involving a specific level of theory.

Qs are the other key parameter obtained from MB spectroscopy arising from the interaction of the quadrupole moment of the first nuclear excited state of ^{57}Fe ($I = 3/2$) with the electric field gradient surrounding the Fe nucleus. QS parameters are very sensitive to the local chemical environment of the Fe nucleus and give information on the population of its 3d orbitals and their ligand environment. The quadrupole interaction splits the ^{57}Fe ($I = 3/2$) nucleus into two doubly degenerate states ($m_I \pm 1/2$ and $m_I \pm 3/2$) whose energy difference is calculated as

$$\Delta E_Q = \frac{1}{2} e Q V_{zz} (1 + \eta^2/3)^{1/2} \quad (2)$$

where e is the positive electric charge, Q is the nuclear quadrupole moment (in barns), and V_{zz} is the electric field gradient. η is an asymmetry parameter defined as $(V_{xx} - V_{yy})/V_{zz}$ with $|V_{zz}| \geq |V_{yy}| \geq |V_{xx}|$.

Density functional theory (DFT) has become an increasingly popular tool for calculating structures and Mössbauer properties.⁴ Its application to Fe-containing and other metalloenzymes to delineate their mechanisms-of-action offers atomic-level understanding of these intricate processes.^{5–7} The success of this approach results from accurate and insightful comparisons with experimental results that allow a rationalization of existing data and the design of new experiments. To further support this synergy between theory and experiment, our aim in the present study is to calibrate Mössbauer parameters for a variety of commonly used DFT functionals including OLYP, OPBE, PW91, BP86, B3LYP, and B3LYP* using a test set comprised of nonheme nitrosyls including low- ($S = 1/2$) and high-spin ($S = 3/2$) $\{\text{FeNO}\}^7$ and $\{\text{Fe}(\text{NO})_2\}^9$ species and other non-nitrosyl iron–sulfur complexes. We follow the nomenclature of Enemark–Feltham to count the electrons in the $\{\text{FeNO}\}^n$ and $\{\text{Fe}(\text{NO})_2\}^n$ species, where n is the number of d electrons on the metal plus the number of electrons in the π^* orbitals of NO.⁸ Nonheme nitrosyls represent an attractive target to probe the relationship between theory and experiment because of their unusual and complex electronic structure.^{9–12} Indeed, the spin densities of iron nitrosyls exhibit broken-symmetry character, and consequently there is a strong tendency for the majority and minority spin densities to be spatially separated.^{13–16} The complexes within our test set possess mono- and polynuclear Fe centers and thus represent a challenge to theory to describe accurately their electronic structure and properties.

A number of DFT calibration studies of Mössbauer isomer shifts and quadrupole splittings have been reported in the literature.^{3,17–26} For example, Neese and co-workers have found that fit parameters are relatively insensitive to whether nonrelativistic or quasi-relativistic (via the zero-order regular approximation, ZORA) DFT methods are employed.^{22,25} The ZORA Fe electron densities are shifted toward the four-component Dirac–Fock limit, but no improvement in the correlation is observed. Further, no improvement in the calculated quadrupole splittings were observed when the ZORA approach was used instead of a nonrelativistic treatment.²² Very recently, Lippard and co-workers reported the performance of eight density

functionals for the prediction of ISs and Qs. They found the B3LYP and O3LYP density functionals to have the lowest errors while M06-2X and SVWN5 possess the highest for their large test set of 31 iron-containing complexes (35 signals), including both mononuclear and dinuclear systems.²⁶

While these previous efforts have proven valuable in their own right, most of the complexes examined have been mononuclear Fe complexes, with the exception of a few. The present effort, which includes mononuclear, dinuclear, and tetranuclear complexes in the training set, is our latest contribution to this growing field and represents an important step forward by providing accurate calibration fits for spectroscopic isomer shifts and quadrupole splitting parameters. One aspect of this work that is particularly valuable is our finding that the commonly used functional B3LYP (and B3LYP*) performs poorly in predicting accurate geometries, especially for the dinuclear and tetranuclear Fe complexes. We find that these poor geometries lead to very poor isomer shift calibration fits.

For the first time in an isomer shift calibration study, we compare two methods used to calculate the Fe electron density. The typical procedure determines this quantity directly at the Fe nucleus. More recently, the ADF software package has by default begun to determine this quantity on a small sphere surrounding the Fe nucleus, which better reflects the underlying physics of the Mössbauer-active nucleus, since the nuclear radius changes from the ground to the excited ^{57}Fe state. Moreover, this approach prohibits the four-component Dirac–Fock electron density from becoming infinite in the limit of a large enough basis set at a point charge nucleus during relativistic calculation. Because it is not known *a priori* how the choice in method used to calculate the Fe electron density will affect isomer shift calibration fits, we sought to examine this issue by calibrating six exchange–correlation potentials for each approach. We believe this work will be widely useful for researchers wishing to calculate accurate Mössbauer parameters for comparison with experimental results.

COMPUTATIONAL METHODS

Geometry optimizations using the PW91,^{27,28} BP86,^{29,30} OLYP,^{31,32} OPBE,^{31,33,34} B3LYP,^{32,35,36} and B3LYP*³⁷ exchange–correlation functionals were carried out with the Amsterdam Density Functional (ADF) 2009.01^{38,39} software package. These particular functionals have been chosen because they have previously been shown to give accurate geometries, energies, and/or spin-state splittings for transition metal complexes with weak field ligands. All-electron Slater-type triple- ζ plus polarization (STO-TZP) basis sets were used for both the geometry optimizations and subsequent Mössbauer calculations, with the numerical integration accuracy set to 5.5. This approach is identical to that taken in our previous work in this area.¹¹ The “AddDiffuseFit” feature of ADF was also employed for the B3LYP and B3LYP* calculations to minimize numerical problems found with hybrid functionals and the fit procedure used by ADF. The ORCA 2.8.0⁴⁰ software package was also used to calculate Mössbauer parameters for the complexes of this study with B3LYP and B3LYP*. For the latter calculations, the Gaussian-type basis set TZVP of Ahlrichs et al.^{41,42} was used for the geometry optimizations with “Grid4”. Single-point energy calculations were then performed using the TZVP basis set for all elements but Fe, where for this element we used the CP(PPP) basis set,³ which utilizes the Ahlrichs (2d2fg, 3p2df) polarization functions from the Turbomole basis set library under ftp.

Table 1. Details of the Iron Complexes Examined in This Study

complex	structure ^a	pt gr.	S_t	Fe oxidation	temp. (K)	δ_{exp} (mm s ⁻¹)	$ \Delta E_{\text{Q}} _{\text{exp}}$ (mm s ⁻¹)	ref ^b
[Fe(SEt) ₄] ⁻	CANDAW10	C ₁	5/2	Fe ³⁺	4.2	0.25	0.62	52, 53
[FeS ₄ C ₈ O ₄] ²⁻	PTSQFE10	C ₁	2	Fe ²⁺	4.2	0.668	3.97	54
[Fe(SPh) ₄] ²⁻	PTHPE10	C ₁	2	Fe ²⁺	4.2	0.66	3.24	54
[Fe(H ₂ O) ₅ (NO)] ²⁺		C ₁	3/2	{FeNO} ⁷	80	0.76	2.10	55
[Fe(NO)(dtc-Pr ₂) ₂] ^c	PRCBFE	C ₁	1/2	{FeNO} ⁷	4.2	0.35	0.89	56, 57
[Fe(SPh) ₂ (NO) ₂] ⁻	SI	C ₁	1/2	{Fe(NO) ₂ } ⁹	4.2	0.182	0.692	58
[Fe(SC ₂ H ₃ N ₃)(SC ₂ H ₂ N ₃)(NO) ₂]	EYABOV	C ₁	1/2	{Fe(NO) ₂ } ⁹	296	0.188	1.118	59, 60
[Fe ₂ S ₂ (S ₂ -o-xyI) ₂] ²⁻	XLDTSF	C _s	0	2*Fe ³⁺	4.2	0.28	0.36	61–63
[Fe ₂ S ₂ (Oph- <i>p</i> -CH ₃) ₄] ²⁻	GIBCUP	C _s	0	2*Fe ³⁺	4.2	0.37	0.32	63
[Fe ₂ S ₂ (C ₄ H ₄ N) ₄] ²⁻	CONSED10	C _{2v}	0	2*Fe ³⁺	77	0.26	0.49	63
[Fe ₂ (NO) ₂ (Et-HPTB)(O ₂ CPh)] ^{2+ d}	RABHAD	C ₁	0	2*{FeNO} ⁷	4.2	0.67	1.44	64
[Fe(NO) ₂ {Fe(NO)(N(CH ₂ CH ₂ S) ₃)}-S,S']	SI	C ₁	1	{FeNO} ⁷	77	0.37	1.15	65
				{Fe(NO) ₂ } ⁹		0.18	1.04	
[Fe ₄ S ₄ (SPh) ₄] ²⁻	FEMJAI02	C ₁	0	4*Fe ^{2.5+}	4.2	0.46	1.07	66, 67
[Fe ₄ S ₄ (Oph) ₄] ²⁻	CAPGAB	C ₁	0	4*Fe ^{2.5+}	4.2	0.50	1.21	68
[Fe ₄ S ₄ (SPh) ₂ Cl ₂] ^{2- g}	CIYKUQ	C ₁	0	2*Fe ^{2.5+} -SPh	4.2	0.48	0.90	66
				2*Fe ^{2.5+} -Cl		0.51	1.22	
[Fe ₄ S ₄ (Oph) ₂ Cl ₂] ^{2- g}	CIYLAX	C ₁	0	2*Fe ^{2.5+} -Oph	4.2	0.51	1.01	66
				2*Fe ^{2.5+} -Cl		0.52	1.28	
[Fe ₄ S ₄ Cl ₄] ²⁻	NUSROI	C _{2v}	0	4*Fe ^{2.5+}	4.2	0.52	1.09	66, 69
[Fe ₄ S ₄ (SCH ₂ CO ₂ Et) ₄] ²⁻	CEQYAY	C ₂	0	4*Fe ^{2.5+}	78	0.43	0.81	70
[Fe ₄ (NO) ₄ (μ ₃ -S) ₄] ^e	KOCBUZ	C _{2v}	0	4*{FeNO} ⁷	78	0.15	1.473	71, 72
[Fe ₄ (NO) ₄ (μ ₃ -S) ₄] ⁻	BIBMOO	C _{2v}	1/2	2*{FeNO} ^{7,5}	RT ^f	0.156	0.935	73
				2*{FeNO} ⁷		0.156	0.935	

^a Cambridge ID or reference to cif file, SI = Supporting Information. ^b References for crystal structures and Mössbauer parameters. ^c dtc = dithiocarbamate. ^d Et-HPTB = *N,N,N',N'*-tetrakis-(*N*-ethyl-2-benzimidazolymethyl)-2-hydroxy-1,3-diaminopropane. ^e Oxidation state assignment dependent upon the electronic state. ^f RT = room temperature, assumed to be 298 K. ^g Experimental values were assigned such that the smaller isomer shift is assigned to the higher electron density.

chemie.unikarlsruhe.de/pub/basen and 2p functions⁴³ plus one additional f function from the Turbomole library. In addition, the “SpecialGrid” option was used to enhance the integration accuracy of Fe to 7.0. In both ADF and ORCA, solvent effects were modeled with the COnductor-like Screening MOdel (COSMO),⁴⁴ using methanol as the solvent (dielectric constant = 32.6).

All of the $S_t = 0, 1/2,$ or 1 polynuclear complexes in this study have high-spin metal sites and exhibit antiferromagnetic coupling. Therefore, spin-unrestricted broken symmetry DFT was used to describe the electronic structure of these systems.⁴⁵ This approach constructs a wave function such that the spin-up (α) electron density occupies a different spatial region than the spin-down (β) electron density. In practice, this is achieved by first constructing a high-spin ferromagnetically coupled state and then exchanging the α and β electron densities of the appropriate iron atoms. Reconverging the wave function then produces the desired antiferromagnetically coupled state.

In some cases, we were not able to converge the geometry of a given species to sufficient accuracy with the B3LYP and B3LYP* exchange-correlation functionals. Consequently, in such cases, we used OLYP geometries for the B3LYP and B3LYP* STO-TZP COSMO single-point energy calculations for the following complexes: [Fe₂S₂(S₂-o-xyI)₂]²⁻, [Fe₄S₄(SCH₂CO₂Et)₄]²⁻, [Fe₄S₄(SPh)₄]²⁻, and [Fe₂(NO)₂(Et-HPTB)(O₂CPh)]²⁺. In addition, we used an OLYP/STO-TZP COSMO geometry for the B3LYP/STO-TZP COSMO single-point energy calculation

for [Fe₄S₄(SPh)₂Cl₂]²⁻. Finally, a B3LYP*/TZVP COSMO geometry was used for the B3LYP/CP(PPP)-TZVP COSMO single-point energy calculation for [Fe₂(NO)₂(Et-HPTB)(O₂CPh)]²⁺.

As previously noted, Mössbauer ISs are proportional to the s-electron density at the nucleus, $\rho(0)$. This quantity is often calculated directly at the nucleus. However, the latest ADF codes (2009 and 2010) calculate the Fe electron density at points on a small spherical surface around the center of the nucleus, where the average electron density over these points is reported.^{1,38,39}

This spherical surface method better represents the underlying physics of the isomer shift phenomena. Upon γ -ray absorption, the ⁵⁷Fe radius changes between the ground and excited states, and the isomer shift is due to the contact interaction of the electron density with the thin spherical shell between these radii. An additional appeal of this approach is to prohibit the four-component Dirac–Fock electron density from becoming infinite in the limit of a large enough basis set at a point charge nucleus during relativistic calculations. The use of a small sphere surrounding the nucleus circumvents this issue because the relativistic electron density remains finite there. Further, if a proper finite-sized nucleus is used, then the relativistic density will remain finite even at the center of the nucleus. In any event, these differences in approach led us to assess how the Mössbauer calibration parameters were affected, since it is valuable to have available in the literature calibration parameters that can be applied to standard output from software packages. To evaluate

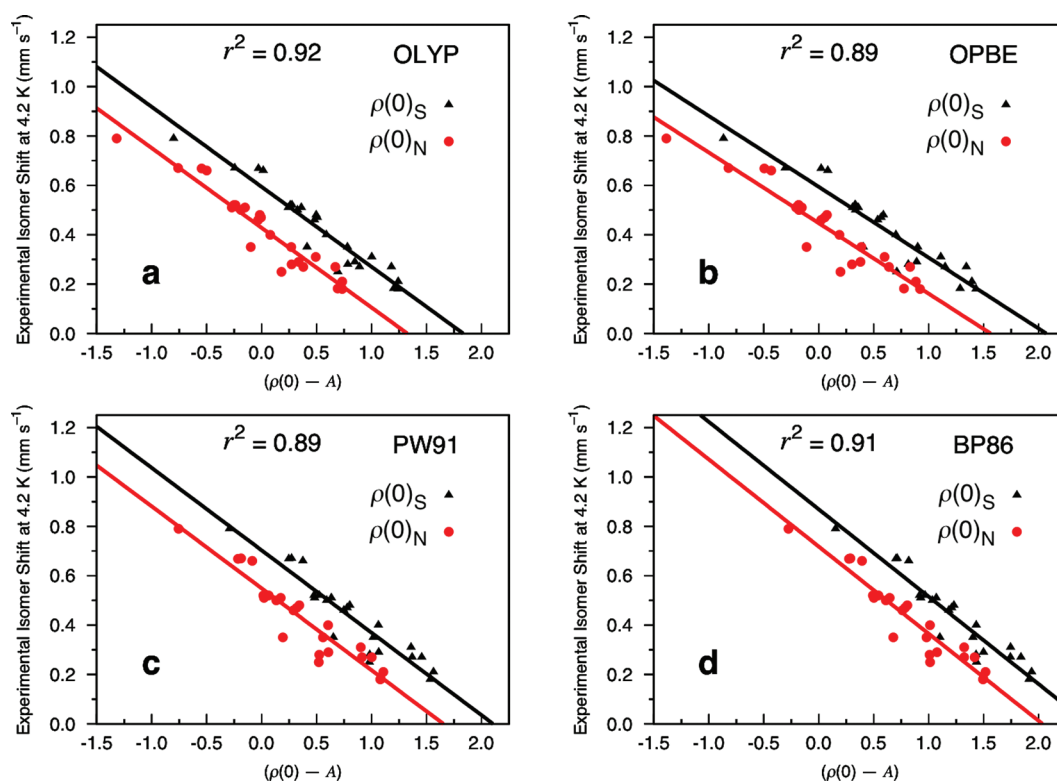


Figure 1. Isomer shift fit based on eight Fe–NO and 12 Fe–S complexes (24 sites) calculated at the OLYP, OPBE, PW91, and BP86 levels of theory using COSMO and the STO-TZP basis set. The Fe electron density is calculated directly at the nucleus ($\rho(0)_N$, circles) and on a small sphere around the center of the Fe nucleus ($\rho(0)_S$, triangles).

the Fe $\rho(0)$ directly at the nucleus in ADF, we use our in-house hyper2003⁴⁶ program in conjunction with ADF's TAPE21, as in our previous work. We note that by default ORCA also evaluates $\rho(0)$ at the nucleus.

Finally, the nuclear quadrupole moment of ^{57}Fe is of interest since ADF and ORCA use slightly different values (0.15 vs 0.16 electron-barns, respectively). Previously, our group had used $eQ = 0.15$ electron-barn.⁴⁷ For the current study, however, we use $eQ = 0.158$ electron-barn, which is taken from recent nonrelativistic quantum chemical calculations.²²

RESULTS AND DISCUSSION

a. Details of the Test Set. Table 1 provides a detailed summary of the complexes used in this study, which follows from our earlier work where the performance of the OLYP exchange-correlation functional to calculate Mössbauer parameters was assessed.¹¹ The 12 non-nitrosyl (Fe–S) complexes of this study include synthetic analogues of the $[\text{Fe}(\text{SR})_4]$ center bacterial rubredoxins, plant-like 2Fe–2S ferredoxins and Rieske proteins, and the 4Fe–4S cubane centers in ferredoxins and high-potential iron proteins (HiPIPs). The three mononuclear Fe–S complexes are high-spin Fe centers, with an S_{tot} of either 5/2 or 2. Table S1 in the Supporting Information shows the Mulliken Fe spin populations of these species calculated at various levels of theory. As is expected for these high-spin Fe^{3+} and Fe^{2+} centers, the magnitudes of these populations range from ca. 3.4 to 3.9, with the Fe^{3+} centers possessing the larger spin populations. All three of the 2Fe–2S complexes possess high-spin $S = 5/2$ Fe^{3+} centers and are antiferromagnetically coupled to give $S_{\text{tot}} = 0$. An examination of the Mulliken Fe spin

populations (Table S1) confirms this, with their (absolute) values ranging from ca. 3.3 to 3.9 for the broken-symmetry states. Similarly, the 4Fe–4S complexes are antiferromagnetically coupled to give $S_{\text{tot}} = 0$, but here each Fe center is considered to possess an oxidation state of +2.5, which may be regarded as resulting from the pairwise interaction of Fe^{2+} ferromagnetically coupled to Fe^{3+} and associated electron delocalization. In these cases, the absolute values of the Fe spin populations range from 2.9 to 3.7. For all of the complexes studied, the hybrid density functionals B3LYP and B3LYP* give higher Mulliken spin populations than do the OLYP, OPBE, PW91, and BP86 density functionals. For the pure functionals, we note that PW91 gives the lowest degree of localized spin on the Fe centers, while OPBE gives the highest.

All eight iron–nitrosyl complexes within this study include $\{\text{FeNO}\}^7$ and/or $\{\text{Fe}(\text{NO})_2\}^9$ moieties. Table S2 (Supporting Information) shows the Fe and NO ligand Mulliken spin populations of each FeNO unit. It can be seen that each unit exhibits oppositely aligned Fe and NO spin populations. Four mononuclear $\{\text{FeNO}\}^7$ and $\{\text{Fe}(\text{NO})_2\}^9$ iron–nitrosyl complexes were examined with $S = 3/2$ or $1/2$. The Fe spin population in the $S = 3/2$ $[\text{Fe}(\text{H}_2\text{O})_5(\text{NO})]^{2+}$ complex ranges from 3.4 to 3.8. Here, the $\{\text{FeNO}\}^7$ unit may be described as a high-spin ($S = 5/2$) Fe^{3+} antiferromagnetically coupled to NO^- ($S = 1$).⁴⁸ In contrast, the FeNO units in $[\text{Fe}(\text{NO})(\text{dtci-Pr}_2)_2]$, $[\text{Fe}(\text{SPh})_2(\text{NO})_2]^-$, and $[\text{Fe}(\text{SC}_2\text{H}_3\text{N}_3)(\text{SC}_2\text{H}_2\text{N}_3)(\text{NO})_2]$ exhibit substantially lower Fe spin populations. For these cases, the $\{\text{FeNO}\}^7$ unit in $[\text{Fe}(\text{NO})(\text{dtci-Pr}_2)_2]$ can be described as an intermediate-spin Fe^{3+} center ($S = 3/2$) coupled to an $S = 1$ NO^- , while the $S = 1/2$ $\{\text{Fe}(\text{NO})_2\}^9$ units in $[\text{Fe}(\text{SPh})_2(\text{NO})_2]^-$ and $[\text{Fe}(\text{SC}_2\text{H}_3\text{N}_3)(\text{SC}_2\text{H}_2\text{N}_3)(\text{NO})_2]$ may be described as high-spin Fe^{3+} antiferromagnetically coupled to two NO^- diradicals or as

high-spin Fe^{1+} ($S = 3/2$) coupled to two NO^* radicals. The interested reader is referred elsewhere for the details of previous DFT calculations and associated molecular orbital arguments to explain the factors that affect the FeNO angle.^{9,49–51}

Two dinuclear nitrosyl complexes included in our study are the diamagnetic bis- $\{\text{FeNO}\}^7$ $[\text{Fe}_2(\text{NO})_2(\text{Et-HPTB})(\text{O}_2\text{CPh})]^{2+}$ and the paramagnetic $S = 1$ thiolate-bridged $\{\text{FeNO}\}^7$ - $\{\text{Fe}(\text{NO})_2\}^9$ complex $[\text{Fe}(\text{NO})_2\{\text{Fe}(\text{NO})(\text{N}(\text{CH}_2\text{CH}_2\text{S})_3)\}-\text{S},\text{S}']$. The dinuclear $S = 0$ $[\text{Fe}_2(\text{NO})_2(\text{Et-HPTB})(\text{O}_2\text{CPh})]^{2+}$ complex contains oppositely aligned $S = 3/2$ $\{\text{FeNO}\}^7$ units with (absolute) spin populations of 3.2–3.8. The $[\text{Fe}(\text{NO})_2\{\text{Fe}(\text{NO})(\text{N}(\text{CH}_2\text{CH}_2\text{S})_3)\}-\text{S},\text{S}']$ complex may be viewed as an $S = 3/2$ $\{\text{FeNO}\}^7$ unit antiferromagnetically coupled with an $S = 1/2$ $\{\text{Fe}(\text{NO})_2\}^9$ unit to give an $S = 1$ complex.

The neutral and reduced $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^{0/-}$ complexes possess 4 Fe –4 S cubane cores, as discussed above. Each iron–NO in the neutral complex is a $\{\text{FeNO}\}^7$ unit, while the reduced form can be described as having two of the iron centers as $\{\text{FeNO}\}^7$ units and two as $\{\text{FeNO}\}^{7.5}$ units. Further details about all of the complexes studied in this work can be found in our previous work.¹¹

b. Isomer Shift. The experimental Fe isomer shifts (IS) used in this study were obtained at temperatures ranging from 4.2 to 298 K. Increases in temperature will result in a lower isomer shift, principally due to the second-order Doppler effect. Thus, to enable a meaningful comparison with our calculated values, we correct the experimental ISs to a common temperature of 4.2 K. This correction is expected to be linear with the temperature⁷⁴ and in the present work is taken to be 0.12 mm s^{-1} over the range 4.2 to 300 K.

Figure 1 displays four of our universal IS fits based on calculations of eight iron–nitrosyl (Fe–NO) and 12 non-nitrosyl (Fe–S) complexes, comprising a total of 24 distinct iron sites. For each level of theory, the Fe electron density is calculated both at the nucleus, $\rho(0)_\text{N}$, and on a small sphere surrounding the nucleus, $\rho(0)_\text{S}$. Generally, the linear regression analysis yields near identical values for α but different values for C and A for the two methods (cf. eq 1).

Figure 1a shows the results for calculations with the OLYP/STO-TZP COSMO level of theory. For $\rho(0)_\text{N}$, the analysis yields values of $\alpha = -0.323$, $C = 0.428$, and $A = 11877$, while for $\rho(0)_\text{S}$, these values are $\alpha = -0.324$, $C = 0.594$, and $A = 11820$. The r^2 value for both methods is 0.92 with a mean absolute error (MAE) of 0.037 mm s^{-1} . The complexes that exhibit the largest MAEs are $[\text{Fe}(\text{SEt})_4]^-$ (0.12 mm s^{-1}) and $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$ (0.09 mm s^{-1}). When these two complexes are removed from the analysis, the total MAE is reduced to 0.029 mm s^{-1} and the r^2 becomes 0.95. The corresponding linear regression values are $\alpha = -0.322$ and $C = 0.603$ for $\rho(0)_\text{S}$ and $\alpha = -0.321$ and $C = 0.437$ for $\rho(0)_\text{N}$ (of course, the A values remain constant). As will be shown below, $[\text{Fe}(\text{SEt})_4]^-$ and $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$ emerge consistently as structures that exhibit the poorest comparisons with experiment for the OLYP, OPBE, PW91, and BP86 functionals. Interestingly, inspection of their calculated geometries does not reveal significant deviations from the experimental structures.

Figure 1b shows our isomer shift fit based on OPBE/STO-TZP COSMO calculations of all of the complexes in our test set. In this case, the linear regression analysis for $\rho(0)_\text{N}$ yields values of $\alpha = -0.286$, $C = 0.447$, and $A = 11877$, while for $\rho(0)_\text{S}$, these values are $\alpha = -0.287$, $C = 0.594$, and $A = 11820$. The r^2 for both methods is 0.89 with a MAE of 0.041 mm s^{-1} , which is slightly

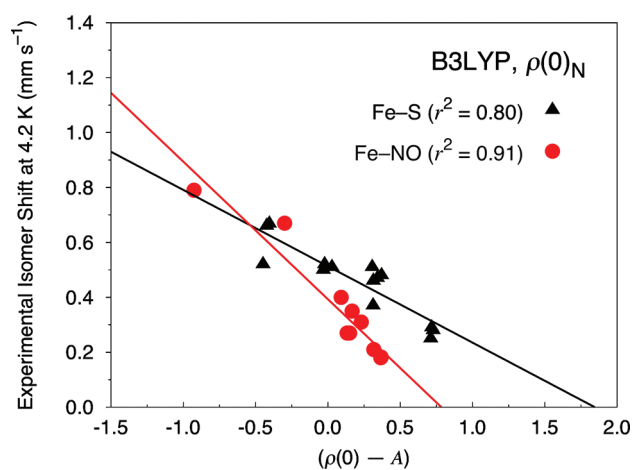


Figure 2. Isomer shift fit based on 12 Fe–S (14 sites, triangles) and eight Fe–NO (10 sites, circles) complexes calculated at the B3LYP/STO-TZP COSMO level of theory with the Fe electron density calculated directly at the nucleus ($\rho(0)_\text{N}$).

worse than is observed for OLYP. As found with OLYP, the calculated isomer shifts of $[\text{Fe}(\text{SEt})_4]^-$ and $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$ show the largest variations relative to experimental results (0.14 and 0.11 mm s^{-1} , respectively), and their removal from the analysis improves the MAE and r^2 appreciably. Specifically, the MAE is reduced to 0.030 mm s^{-1} with an r^2 of 0.94. The linear regression parameters then become $\alpha = -0.287$ and $C = 0.458$ for $\rho(0)_\text{N}$ and $\alpha = -0.288$ and $C = 0.606$ for $\rho(0)_\text{S}$.

Figure 1c displays the results we obtain with our PW91/STO-TZP COSMO calculations. Overall, PW91 yields a slightly worse MAE (of 0.042 mm s^{-1}) relative to experimental results for our complete test set than is observed for either OLYP or OPBE. The linear regression analysis for $\rho(0)_\text{N}$ yields values of $\alpha = -0.332$, $C = 0.549$, and $A = 11874$. These values are $\alpha = -0.334$, $C = 0.703$, and $A = 11827$ for $\rho(0)_\text{S}$, with an r^2 value of 0.89. $[\text{Fe}(\text{SEt})_4]^-$ and $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$ again show the largest variations relative to experimental results (0.13 and 0.12 mm s^{-1} , respectively). Removing them from the analysis yields a MAE of 0.034 mm s^{-1} and an r^2 of 0.94. We also observe slightly changed linear regression values for $\rho(0)_\text{N}$ and $\rho(0)_\text{S}$ of $\alpha = -0.333$ and $C = 0.561$ and $\alpha = -0.335$ and $C = 0.715$, respectively.

Interestingly, BP86 performs quite well in our universal calibration analysis (Figure 1d), yielding a total MAE of 0.040 mm s^{-1} and an r^2 of 0.91 for the test set, which places it second to the best performing functional thus far (OLYP and MAE of 0.037 mm s^{-1} and r^2 of 0.92). Calculation of the Fe electron density at the nucleus, $\rho(0)_\text{N}$, yields linear regression values of $\alpha = -0.353$, $C = 0.718$, and $A = 11889$, while calculation of the density at points on a small sphere around the center of the Fe nucleus, $\rho(0)_\text{S}$, gives values of $\alpha = -0.354$, $C = 0.869$, and $A = 11832$. Consistent with our previous findings, $[\text{Fe}(\text{SEt})_4]^-$ and $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$ give the largest variations in calculated δ relative to experimental results (0.11 mm s^{-1} for each). The MAE improves to 0.033 mm s^{-1} when these two structures are removed from the analysis, with an r^2 of 0.95. The corresponding linear regression values become $\alpha = -0.353$ and $C = 0.728$ for $\rho(0)_\text{N}$ and $\alpha = -0.354$ and $C = 0.879$ for $\rho(0)_\text{S}$.

In Figure S1 in the Supporting Information, we show the universal calibration fitting results obtained with the B3LYP/STO-TZP COSMO level of theory. Overall, we find a

Table 2. Comparison of Deviations from Experimental Results and Mean Absolute Deviations (MAD) of Selected Calculated Average Bond Distances (Å) of the Fe₄S₄ Structures of This Study^a

		exptl.	OLYP	OPBE	PW91	BP86	B3LYP	B3LYP*	B3LYP ^b	B3LYP ^{*b}
[Fe ₄ S ₄ (SPh) ₄] ²⁻	Fe–Fe	2.736	−0.012	−0.075	−0.066	−0.052	n/a ^c	n/a ^c	+0.144	+0.147
	Fe–S	2.286	+0.022	−0.017	−0.014	−0.007			+0.071	+0.075
[Fe ₄ S ₄ (OPh) ₄] ²⁻	Fe–Fe	2.753	+0.017	−0.031	−0.051	−0.037	+0.205	+0.157	+0.150	+0.153
	Fe–S	2.294	+0.028	−0.010	−0.008	−0.002	+0.076	+0.062	+0.071	+0.076
[Fe ₄ S ₄ (SPh) ₂ Cl ₂] ²⁻	Fe–Fe	2.745	+0.010	−0.071	−0.060	−0.043	n/a ^c	+0.127	+0.155	+0.110
	Fe–S	2.278	+0.029	−0.011	−0.007	−0.001		+0.066	+0.080	+0.033
[Fe ₄ S ₄ (OPh) ₂ Cl ₂] ²⁻	Fe–Fe	2.766	+0.005	−0.061	−0.060	−0.001	+0.206	+0.133	+0.159	+0.155
	Fe–S	2.285	+0.028	−0.007	−0.005	+0.001	+0.084	+0.063	+0.079	+0.080
[Fe ₄ S ₄ Cl ₄] ²⁻	Fe–Fe	2.700	+0.005	−0.072	−0.063	−0.049	+0.159	+0.114	+0.130	+0.126
	Fe–S	2.289	+0.017	+0.017	−0.014	−0.009	+0.064	+0.051	+0.063	+0.066
[Fe ₄ S ₄ (SCH ₂ CO ₂ Et) ₄] ²⁻	Fe–Fe	2.754	−0.008	−0.093	−0.084	−0.066	n/a ^c	n/a ^c	+0.116	+0.117
	Fe–S	2.303	+0.018	−0.015	−0.015	−0.013			+0.098	+0.077
[Fe ₄ (NO) ₄ (μ ₃ -S) ₄]	Fe–Fe	2.651	−0.016	−0.097	−0.049	−0.045	+0.208	+0.131	+0.307	+0.298
	Fe–S	2.217	−0.010	−0.064	−0.038	−0.034	+0.094	+0.069	+0.143	+0.145
[Fe ₄ (NO) ₄ (μ ₃ -S) ₄] ⁻	Fe–Fe	2.688	+0.015	−0.080	−0.039	−0.031	+0.208	+0.151	+0.232	+0.229
	Fe–S	2.231	+0.008	−0.051	−0.032	−0.027	+0.107	+0.083	+0.127	+0.129
MAD (Fe–Fe)			0.011	0.073	0.059	0.041	0.197	0.136	0.174	0.167
MAD (Fe–S)			0.020	0.023	0.017	0.012	0.085	0.066	0.092	0.085

^a Calculated values reported as deviations from experimental results. ^b Calculated with ORCA using COSMO in combination with a CP(PPP) basis set on Fe and a TZVP basis set on the remaining atoms. ^c Geometry is not available at this level of theory. The OLYP/STO-TZP COSMO geometry is used instead. See the Computational Methods section for details.

poor linear correlation of the calculated Fe electron density versus the experimental IS for our test set as the linear regression analysis gives an r^2 value of 0.64 with a MAE of 0.086 mm s^{−1}. This poor performance is rather surprising since several other studies show B3LYP to demonstrate excellent performance.^{22,25,26} This discrepancy appears to be both a consequence of the makeup of our test set and the relatively poor geometries we obtain with this functional for polynuclear Fe complexes (see below).

Previously, it was shown that grouping different oxidation states of Fe will produce different IS calibration lines.²⁴ Hopmann et al. recently observed slightly improved linear regression parameters when the Fe–NO complexes of their test set were determined separately.¹¹ Specifically, the fit r^2 value is reported to change from 0.915 to 0.979, while the MAE decreased from 0.039 to 0.029 mm s^{−1} for a test set that included only Fe–NO complexes.

In this spirit, when our current test set is split into non-nitrosyl (i.e., Fe–S) or iron–nitrosyl (i.e., Fe–NO) complexes, we observe a very good correlation between the calculated $\rho(0)_N$ and the experimental isomer shift for the Fe–NO complexes with the B3LYP functional (Figure 2), but not for the Fe–S complexes. Similar results for when the Fe electron density is calculated on a small sphere around the center of the Fe nucleus can be found in Figure S2 in the Supporting Information. For completeness, we note also that Figures S3–S10 of the Supporting Information correspond to linear correlation plots for the OLYP, OPBE, PW91, and BP86 levels of theory when the test set is split into Fe–S and Fe–NO complexes.

When only the 12 Fe–S complexes are considered, which include 14 distinct Fe sites, we obtain linear correlation parameters for $\rho(0)_N$ of $\alpha = -0.278$, $C = 0.513$, and $A = 11880$, with an r^2 value of 0.80 and a MAE of 0.045 mm s^{−1}. The largest IS outliers in this set correspond to the Fe–Cl and Fe–OPh sites in [Fe₄S₄(OPh)₂Cl₂]²⁻, with MAEs of 0.12 and 0.08 mm s^{−1},

respectively. It is interesting to note that when this complex is removed from our analysis the r^2 value improves greatly to 0.92.

Turning our attention to the eight Fe–NO complexes, we find an r^2 of 0.91; a MAE of 0.050 mm s^{−1}; and linear correlation parameters for $\rho(0)_N$ of $\alpha = -0.501$, $C = 0.393$, and $A = 11880$. With a MAE of 0.13 mm s^{−1}, the {FeNO}⁷ site in [Fe₂(NO)₂-(Et-HPTB)(O₂CPh)]²⁺ is the largest outlier. We note that an OLYP geometry was used for this structure because we were unable to converge the B3LYP geometry to sufficient accuracy. However, it is unlikely that this approximation is the primary source of this error, since in many cases excellent agreement is observed between our calculated OLYP structures and those obtained from experimental results (see, e.g., Tables 2 and 3).

As found with our B3LYP calculations, B3LYP* performs worse in predicting accurate ISs than OLYP, OPBE, PW91 or BP86. Unlike B3LYP (Figure S1, Supporting Information), however, B3LYP* is found to show reasonable performance over the entire test set (Figure 3). The total MAE over the entire test set is 0.053 mm s^{−1} with an r^2 of 0.86. The linear regression values for the calculation of ISs at this level of theory for $\rho(0)_N$ are $\alpha = -0.401$, $C = 0.804$, and $A = 11876$. When the Fe electron density is calculated as $\rho(0)_S$, the values are $\alpha = -0.402$, $C = 0.610$, and $A = 11820$. The Fe sites that yield the highest MAEs (of 0.10 mm s^{−1} each) relative to the experimental results for B3LYP* include [Fe₄S₄(SCH₂CO₂Et)₄]²⁻, the {FeNO}^{7,5} fragment of [Fe₄(NO)₄(μ₃-S)₄]⁻, and the Fe–SPh sites in [Fe₄S₄(SPh)₂Cl₂]²⁻. Interestingly, the poor performance of OLYP, OPBE, PW91, and BP86 in predicting accurate isomer shifts for [Fe(SET)₄]⁻ and [Fe₂S₂(OPh-*p*-CH₃)₄]²⁻ is not found with B3LYP* (or B3LYP for that matter). In fact, two of the lowest MAEs determined from these B3LYP* calculations are [Fe(SET)₄]⁻ (0.02 mm s^{−1}) and [Fe₂S₂(OPH-*p*-CH₃)₄]²⁻ (0.01 mm s^{−1}).

Improved performance of B3LYP* is observed if our test set is split into separate groupings of non-nitrosyl (Fe–S) and

Table 3. Comparison of Deviations from Experimental Results of Selected Calculated Average Bond Distances (Å) and Angles (deg) of Selected Mononuclear Fe Structures of This Study^a

		exptl.	OLYP	OPBE	PW91	BP86	B3LYP	B3LYP*	B3LYP ^b	B3LYP ^{*b}
[Fe(SET) ₄] ⁻	Fe–S	2.269	+0.061	+0.028	+0.029	+0.032	+0.052	+0.048	+0.055	+0.068
	S–C	1.880	–0.033	–0.051	–0.029	–0.023	–0.026	–0.025	–0.017	–0.010
[Fe(SPh) ₄] ²⁻	Fe–S	2.353	–0.012	–0.053	–0.053	–0.047	+0.046	+0.026	+0.046	+0.059
	S–C	1.767	+0.006	–0.008	+0.011	+0.015	+0.021	+0.014	+0.021	+0.025
[FeS ₄ C ₈ O ₄] ²⁻	Fe–S	2.389	+0.011	–0.030	–0.014	–0.007	+0.060	+0.045	+0.051	+0.063
	S–C	1.688	+0.011	0.000	+0.014	+0.017	+0.016	+0.016	+0.027	+0.030
[Fe(NO)(dtci-Pr ₂) ₂]	Fe–S	2.288	+0.016	–0.031	+0.006	+0.015	+0.056	+0.045	+0.060	+0.068
	Fe–N	1.676	–0.008	–0.034	–0.017	–0.015	+0.083	+0.040	+0.083	+0.060
	N–O	1.161	+0.032	+0.021	+0.033	+0.034	+0.027	+0.024	+0.018	+0.017
[Fe(SC ₂ H ₃ N ₃)(SC ₂ H ₂ N ₃)(NO) ₂]	Fe–S	2.308	+0.039	–0.011	–0.010	–0.005	+0.084	+0.065	+0.081	+0.092
	Fe–N	1.673	–0.002	–0.023	–0.012	–0.010	+0.083	+0.054	+0.085	+0.071
	N–O	1.154	+0.028	+0.021	+0.029	+0.031	+0.023	+0.022	+0.016	+0.017
MAD (Fe–S)			0.028	0.030	0.022	0.021	0.060	0.048	0.058	0.070
MAD (S–C)			0.016	0.020	0.017	0.019	0.022	0.020	0.023	0.022
MAD (Fe–N)			0.006	0.027	0.016	0.015	0.082	0.047	0.083	0.066
MAD (N–O)			0.029	0.021	0.030	0.031	0.027	0.021	0.018	0.019

^a Calculated values reported as deviations from experimental results. ^b Calculated with ORCA using COSMO in combination with a CP(PPP) basis set on Fe and a TZVP basis set on the remaining atoms.

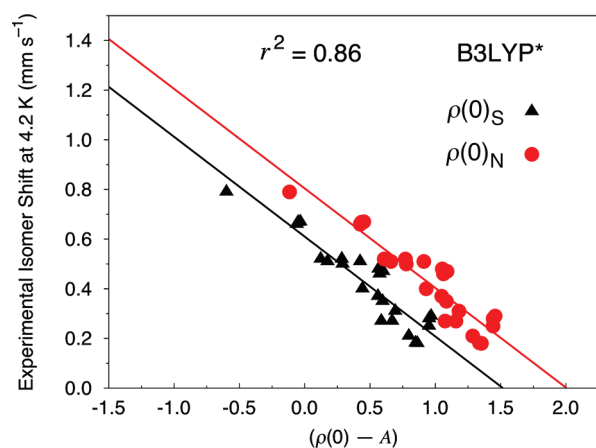


Figure 3. Isomer shift fit based on eight Fe–NO and 12 Fe–S complexes (24 sites) calculated at the B3LYP*/STO-TZVP COSMO level of theory. The Fe electron density is calculated directly at the nucleus ($\rho(0)_N$, circles) and on a small sphere around the center of the Fe nucleus ($\rho(0)_S$, triangles).

iron–nitrosyl (Fe–NO) complexes. Figure 4 shows the results when the Fe electron density is calculated at the nucleus, $\rho(0)_N$; see Figure S11 in the Supporting Information for the analogous data for $\rho(0)_S$. If only the 12 non-nitrosyl complexes are considered, an r^2 value of 0.90 and a MAE of just 0.034 mm s^{-1} is determined for B3LYP*. In this case, all of the complexes exhibit very good MAEs of not more than 0.06 mm s^{-1} . For the eight iron–nitrosyl complexes, a MAE of 0.033 mm s^{-1} with an r^2 of 0.96 is found. Similar to our observations of B3LYP, the largest outlier within the iron–nitrosyl test set is the {FeNO}⁷ site in [Fe₂(NO)₂(Et-HPTB)(O₂CPh)]²⁺, which here possesses a MAE of 0.08 mm s^{-1} .

We extended the scope of the B3LYP and B3LYP* calculations through the use of basis sets derived from Gaussian-type orbitals

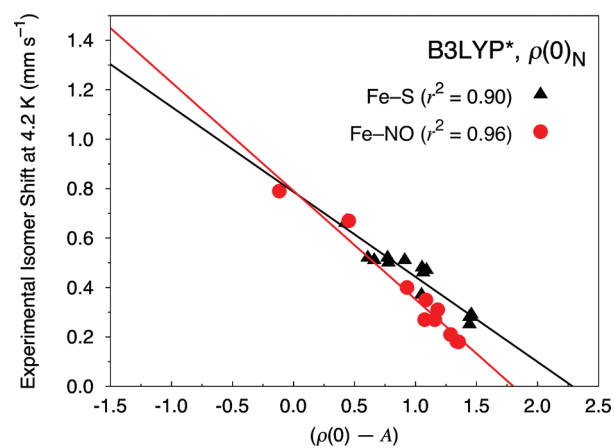


Figure 4. Isomer shift fit based on 12 Fe–S complexes (14 sites, triangles) and eight Fe–NO complexes (10 sites, circles) calculated at the B3LYP*/STO-TZVP COSMO level of theory with the Fe electron density calculated directly at the nucleus ($\rho(0)_N$).

(GTOs, refer to the Computational Methods section for further details). In this case, we find that B3LYP/CP(PPP)-TZVP and B3LYP*/CP(PPP)-TZVP yield similar trends to our results obtained with Slater-type orbitals. That is, B3LYP* performs slightly better than B3LYP, but both methods show limitations in providing good quality universal fit parameters. Indeed, a universal fit for B3LYP using GTOs gives an r^2 of 0.69 and a MAE of 0.079 mm s^{-1} . For B3LYP*, these quantities are 0.77 and 0.064 mm s^{-1} , respectively. Clearly, the rather poor performance of these two functionals over the entire test set appears to be independent of the type of basis set used.

Figures S12 and S13 in the Supporting Information show the calibration lines when we separate the non-nitrosyl (Fe–S) and iron–nitrosyl (Fe–NO) complexes for our B3LYP and B3LYP* calculations that use the CP(PPP)-TZVP basis set. The B3LYP

Table 4. Comparison of Deviations from Experimental Results of Selected Calculated Average Bond Distances (Å) and Angles (deg) of Selected Dinuclear Fe Structures of This Study^a

		exptl.	OLYP	OPBE	PW91	BP86	B3LYP	B3LYP*	B3LYP ^b	B3LYP ^{*b}
[Fe ₂ S ₂ (S ₂ -o-xy ^l) ₂] ²⁻	Fe–Fe	2.698	+0.018	−0.032	−0.021	−0.006	n/a ^c	n/a ^c	+0.143	+0.142
	Fe–S	2.257	+0.026	−0.008	−0.002	+0.004			+0.049	+0.058
	Fe–S–Fe	75.3	−0.4	−0.8	−0.6	−0.4			+2.1	+1.9
[Fe ₂ S ₂ (OPh- <i>p</i> -CH ₃) ₄] ²⁻	Fe–Fe	2.749	+0.049	+0.011	−0.032	−0.027	+0.172	+0.122	+0.126	+0.125
	Fe–S	2.223	+0.031	+0.004	+0.001	+0.004	+0.068	+0.058	+0.061	+0.064
	Fe–S–Fe	76.4	+0.4	+0.2	−1.1	−1.0	+2.8	+1.7	+1.6	+1.5
[Fe ₂ S ₂ (C ₄ H ₄ N) ₄] ²⁻	Fe–Fe	2.677	+0.086	+0.030	−0.003	+0.016	+0.177	+0.142	+0.163	+0.167
	Fe–S	2.191	+0.037	+0.007	+0.006	+0.012	+0.068	+0.057	+0.071	+0.074
	Fe–S–Fe	75.1	+1.5	+0.9	−0.1	+0.3	+3.3	+2.6	+2.7	+2.7
[Fe(NO) ₂ {Fe(NO)(N(CH ₂ CH ₂ S) ₃)} ₂ -S,S']	Fe–Fe	2.766	+0.051	−0.055	−0.049	−0.037	+0.282	+0.224	+0.254	+0.257
	Fe–S	2.297	+0.029	−0.025	−0.022	−0.013	+0.103	+0.084	+0.099	+0.105
	Fe–S–Fe	73.4	+0.6	−0.6	−0.6	−0.5	+4.7	+3.7	+3.9	+3.8
MAD (Fe–Fe)			0.051	0.032	0.026	0.022	0.158	0.122	0.172	0.173
MAD (Fe–S)			0.031	0.011	0.008	0.008	0.080	0.066	0.070	0.075
MAD (Fe–S–Fe)			0.7	0.6	0.6	0.6	3.6	2.7	2.6	2.5

^a Calculated values reported as deviations from experimental results. ^b Calculated with ORCA using COSMO in combination with a CP(PPP) basis set on Fe and a TZVP basis set on the remaining atoms. ^c Geometry is not available at this level of theory. The OLYP/STO-TZP COSMO geometry is used instead. See the Computational Methods section for details.

linear regression parameters for the Fe–S complexes are $\alpha = -0.285$, $C = 0.641$, and $A = 11\,816$. Relative to universal fitting parameters, the r^2 value improves from 0.69 to 0.82 and the MAE reduces from 0.079 to 0.044 mm s^{−1} when the Fe–S complexes are considered separately. The largest outliers in this case are [Fe₄S₄(SPh)₂Cl₂]²⁻ and [Fe₄S₄(OPh)₂Cl₂]²⁻, with MAEs of 0.08 or 0.09 mm s^{−1} for the Fe sites. The Fe–NO complexes give calibration parameters of $\alpha = -0.476$, $C = 0.614$, and $A = 11\,816$. The MAE is 0.048 mm s^{−1}, and the r^2 value improves significantly to 0.93. The largest outlier within the Fe–NO complex test set (with a MAE of 0.09 mm s^{−1}) is the {FeNO}⁷ site in [Fe₄(NO)₄(μ_3 -S)₄]⁻.

Figure S13 (Supporting Information) shows results for our B3LYP*/CP(PPP)-TZVP COSMO calculations. Here, the B3LYP* calibration line for the iron–sulfur complexes yields an r^2 value of 0.84 and an overall MAE of 0.037 mm s^{−1}. The linear regression parameters are $\alpha = -0.326$, $C = 0.529$, and $A = 11\,815$. The largest outlier in this fit is for the Fe–SPh site in [Fe₄S₄(SPh)₂Cl₂]²⁻, which gives an MAE relative to the experimental results of 0.12 mm s^{−1}. The Fe–NO complexes exhibit linear regression parameters of $\alpha = -0.435$, $C = 0.428$, and $A = 11\,815$, with an r^2 of 0.93 and an overall MAE of 0.045 mm s^{−1}. The largest outliers within the Fe–NO complexes are for the {FeNO}⁷ sites in [Fe₄(NO)₄(μ_3 -S)₄] as well as [Fe₂(NO)₂(Et-HPTB)(O₂C-Ph)]²⁺, which each possess MAEs of 0.08 mm s^{−1}.

Overall, we note that regardless of the type of basis set used, B3LYP and B3LYP* give structures that deviate more from the experimental results than is observed with structures calculated with the OLYP, OPBE, PW91, and BP86 functionals (Tables 2, 3, and 4). This is especially true for the Fe–Fe distances in the 4Fe–4S polynuclear complexes of our test set, where we observe mean absolute deviations of 0.136–0.197 Å for B3LYP and B3LYP* relative to the experimental structures (Table 2). In contrast, the Fe–Fe distances calculated with OLYP, OPBE, PW91, and BP86 deviate from experiment no more than 0.1 Å and possess MADs from 0.011 to 0.073 Å. Likewise, B3LYP and B3LYP* also overestimate the Fe–S bond distances, with MADs

ranging from 0.066 to 0.092 Å, which are significantly larger than the MADs observed for the other functionals (0.012 to 0.023 Å, Table 2).

The relatively poor performance of B3LYP and B3LYP* in predicting accurate geometries is also evident in the mono- and dinuclear Fe complexes. For example, Table 3 shows the Fe–S bond distances in the mononuclear Fe complexes to deviate from experimental results slightly more for these hybrid density functionals (MADs ranging from 0.048 to 0.070 Å) than for OLYP (MAD of 0.028 Å), OPBE (MAD of 0.030 Å), PW91 (MAD of 0.022 Å), and BP86 (MAD of 0.021 Å). Similar behavior is observed for the Fe–N bond distances. However, for both the S–C and the N–O distances in the mononuclear complexes, all of the functionals tend to perform with comparable accuracy.

For the four dinuclear Fe complexes in Table 4, we note the very large MADs for the Fe–Fe distance for B3LYP and B3LYP* (ranging from 0.122 to 0.173 Å) compared with the other functionals (MADs ranging from 0.022 to 0.051 Å). Further, we also observe large MADs from experimental results for the Fe–S distances (ranging from 0.060 to 0.080 Å), as well as for the Fe–S–Fe bridge angle (ranging from 2.5° to 3.6°) for the B3LYP and B3LYP* functionals. These relatively large MADs are not observed with the results obtained with OLYP, OPBE, PW91, or BP86 (Table 4). The observation that pure functionals give geometries in better agreement with experimental results than hybrid functionals like UB3LYP has also been found for a series of iron–nitrosyl complexes, including Roussin's red and black salts.¹⁶ Further, significant variations in Fe–NO spin densities calculated with different exchange-correlation functionals were observed.^{12,15,16,75,76}

The possible role of dispersion to correct the geometries obtained with B3LYP (and B3LYP*) has been examined by calculating a representative set of structures from our test set with the B3LYP-D functional, which includes a dispersion correction.⁷⁷ Table 5 reports the deviations from the experimental structures for four mononuclear Fe complexes, two

Table 5. Comparison of Deviations from Experiment of Selected Calculated Average Bond Distances (Å) and Angles (deg) of Selected Mono- and Polynuclear Fe Structures of This Study^a

		OLYP	B3LYP	B3LYP-D	experiment
[Fe(SEt) ₄] ⁻	Fe–S	+0.061	+0.052	+0.042	2.269
	S–C	–0.033	–0.026	–0.028	1.880
[FeS ₄ C ₈ O ₄] ²⁻	Fe–S	+0.011	+0.060	+0.055	2.389
	S–C	+0.011	+0.016	+0.018	1.688
[Fe(NO)(dtci-Pr ₂) ₂]	Fe–S	+0.016	+0.056	+0.054	2.288
	Fe–N	–0.008	+0.083	+0.081	1.676
	N–O	+0.032	+0.027	+0.026	1.161
[Fe(SC ₂ H ₃ N ₃)(SC ₂ H ₂ N ₃)(NO) ₂]	Fe–S	+0.039	+0.084	+0.068	2.308
	Fe–N	–0.002	+0.083	+0.079	1.673
	N–O	+0.028	+0.023	+0.023	1.154
[Fe ₂ S ₂ (C ₄ H ₄ N) ₄] ²⁻	Fe–Fe	+0.086	+0.177	+0.157	2.677
	Fe–S	+0.037	+0.068	+0.063	2.191
	Fe–S–Fe	+1.5	+3.3	+2.8	75.1
[Fe(NO) ₂ {Fe(NO)(N(CH ₂ CH ₂ S) ₃)} ₂ -S,S']	Fe–Fe	+0.051	+0.282	+0.269	2.766
	Fe–S	+0.029	+0.103	+0.095	2.297
	Fe–S–Fe	+0.6	+4.7	+4.7	73.4
[Fe ₄ S ₄ Cl ₄] ²⁻	Fe–Fe	+0.005	+0.159	+0.216	2.700
	Fe–S	+0.017	+0.064	+0.070	2.289
[Fe ₄ (NO) ₄ (μ ₃ -S) ₄]	Fe–Fe	–0.016	+0.208	+0.245	2.651
	Fe–S	–0.010	+0.094	+0.101	2.217
[Fe ₄ (NO) ₄ (μ ₃ -S) ₄] ⁻	Fe–Fe	+0.015	+0.208	+0.255	2.688
	Fe–S	+0.008	+0.107	+0.112	2.231

^a OLYP, B3LYP, and B3LYP-D values use the STO-TZP basis set with a COSMO representation of the solvent. See the Computational Methods for further details.

dinuclear Fe complexes, and three tetranuclear Fe complexes. In general, B3LYP-D reduces the deviations from experimental results for the mono- and dinuclear complexes relative to structures obtained with B3LYP. However, the geometries obtained with B3LYP-D for the tetranuclear complexes show even larger deviations than we observe for geometries calculated with B3LYP. Also, reproduced in Table 5 are the deviations from experimental results for the geometries calculated with the OLYP functional, which show OLYP's rather good performance across all types of Fe complexes.

It is also instructive to consider spin expectation values of our complexes and the possible role of spin contamination and its effect on the calculated geometries. Spin contamination arises from the mixing of higher spin states into a given wave function. Previously, it has been shown that broken symmetry (BS) state geometries exhibit longer Mn–Mn and Fe–Fe distances in [Mn(III)₂(μ-O)₃(NH₃)₆]²⁻ and [Fe₂S₂(SCH₃)₄]²⁻ than their spin-projected *S* = 0 ground state geometries.⁷⁸ Similar results have also been found for the Cr dimer.⁷⁹ It is therefore reasonable to conceive that spin contamination may adversely affect the geometries of the Fe complexes of this study.

Table 6 reports the pure spin-state expectation values, the broken-symmetry (BS) spin-state expectation values in the weak-coupling (WC) limit, and the BS spin-state expectation values of our complexes calculated at various levels of theory. For the iron–nitrosyl complexes, we assume that the coupling within the Fe–NO unit is strong, such that it can be described as being essentially covalent. The pure spin-state values are obtained in the usual way via the relation *S*(*S* + 1), where *S* is the

total spin. The calculated BS values are obtained using eq 3⁸⁰ or 4⁸¹

$$\langle \hat{S}^2 \rangle_{\text{BS}} = \left(\frac{N^\alpha - N^\beta}{2} \right) \left(\frac{N^\alpha - N^\beta}{2} + 1 \right) + N^\beta - \sum_i \sum_j^N |S_{ij}^{\alpha\beta}|^2 \quad (3)$$

$$\langle \hat{S}^2 \rangle_{\text{BS(cot)}} = \left(\frac{N^\alpha - N^\beta}{2} \right) \left(\frac{N^\alpha - N^\beta}{2} + 1 \right) + N^\beta - \sum_i n_i^\alpha n_i^\beta |\tilde{S}_i^{\alpha\beta}|^2 \quad (4)$$

where *N*^α and *N*^β are the number of spin-up and spin-down electrons, respectively, *S*_{*ij*} are the overlap integrals, and *n*_{*i*}^α and *n*_{*i*}^β are the spin–orbital occupation numbers. Equation 3 is the standard method used to determine expectation values from an unrestricted single-determinant wave function,⁸⁰ while eq 4 pertains to the magnetic orbital pairs (with overlap squared |*S*_{*ii*}^{αβ}|²) obtained via the corresponding orbital transformation (cot).⁸¹ Further, *N*^α and *N*^β can be defined as

$$N^\alpha = N_W^\alpha + N_D \quad N^\beta = N_W^\beta + N_D \quad (5)$$

where *N*_W^α and *N*_W^β are the number of (*W* for weakly coupled) spin-up and spin-down electrons in singly occupied orbitals, respectively, and 2*N*_D is the total number of electrons in doubly occupied orbitals. The contributions from the singly and doubly

Table 6. Pure and Broken Symmetry Spin-State Expectation Values for the Iron Complexes Examined in This Study^a

complex	Fe oxidation	$\langle S^2 \rangle_{\text{pure}}$	$\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$	OLYP	OPBE	PW91	BP86	B3LYP		B3LYP*	
								//B3LYP	//OLYP	//B3LYP*	//OLYP
[Fe(SEt) ₄] ⁻	Fe ³⁺	8.75	n/a	8.76	8.76	8.76	8.76	8.77	8.77	8.77	8.77
[FeS ₄ C ₈ O ₄] ²⁻	Fe ²⁺	6.00	n/a	6.06	6.07	6.02	6.02	6.01	6.01	6.01	6.01
[Fe(SPh) ₄] ²⁻	Fe ²⁺	6.00	n/a	6.03	6.03	6.02	6.02	6.02	6.00	6.02	6.02
[Fe(H ₂ O) ₅ (NO)] ²⁺	{FeNO} ⁷	3.75	5.75	4.28	4.29	4.12	4.14	4.53	4.49	4.45	4.41
[Fe(NO)(dtci-Pr ₂) ₂]	{FeNO} ⁷	0.75	2.75	0.86	0.83	0.79	0.79	1.74	1.37	1.36	1.13
[Fe(SPh) ₂ (NO) ₂] ⁻	{Fe(NO) ₂ } ⁹	0.75	4.75	1.02	0.95	0.82	0.83	2.65	2.07	2.12	1.67
[Fe(SC ₂ H ₃ N ₃)(SC ₂ H ₂ N ₃)(NO) ₂]	{Fe(NO) ₂ } ⁹	0.75	4.75	1.08	0.99	0.84	0.86	2.54	2.08	2.08	1.71
MAD ($\langle S^2 \rangle_{\text{pure}}$)				0.19	0.17	0.09	0.10	0.79	0.58	0.58	0.43
[Fe ₂ S ₂ (S ₂ -o-xyI) ₂] ²⁻	2*Fe ³⁺	0.0	5.00	4.31	4.27	4.04	4.08	n/a ^c	4.65	n/a ^c	4.55
[Fe ₂ S ₂ (OPh- <i>p</i> -CH ₃) ₄] ²⁻	2*Fe ³⁺	0.0	5.00	4.47	4.45	4.21	4.22	4.80	4.74	4.72	4.67
[Fe ₂ S ₂ (C ₄ H ₄ N) ₄] ²⁻	2*Fe ³⁺	0.0	5.00	4.31	4.27	3.99	4.03	4.69	4.64	4.59	4.55
[Fe ₂ (NO) ₂ (Et-HPTB)(O ₂ CPh)] ²⁺	2*{FeNO} ⁷	0.0	3.00	4.24	4.20	3.62	3.67	n/a ^c	4.90	n/a ^c	4.69
[Fe(NO) ₂ {Fe(NO)(N(CH ₂ CH ₂ S) ₃)-S,S'}]	{FeNO} ⁷ {Fe(NO) ₂ } ⁹	2.00	3.00	3.06	2.76	2.32	2.38	5.54	4.72	4.83	4.11
MAD ($\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$)				0.64	0.69	0.81	0.79	1.02	0.92	0.84	1.01
[Fe ₄ S ₄ (SPh) ₄] ²⁻	4*Fe ^{2.5+}	0.0	9.00	7.26	7.02	6.23	6.36	n/a ^c	8.24	n/a ^c	7.99
[Fe ₄ S ₄ (OPh) ₄] ²⁻	4*Fe ^{2.5+}	0.0	9.00	7.59	7.42	6.71	6.81	8.67	8.43	8.45	8.24
[Fe ₄ S ₄ (SPh) ₂ Cl ₂] ²⁻	4*Fe ^{2.5+}	0.0	9.00	7.35	7.11	6.37	6.47	8.47	8.28	8.29	8.05
[Fe ₄ S ₄ (OPh) ₂ Cl ₂] ²⁻	4*Fe ^{2.5+}	0.0	9.00	7.49	7.34	6.64	6.75	8.61	8.37	8.40	8.16
[Fe ₄ S ₄ Cl ₄] ²⁻	4*Fe ^{2.5+}	0.0	9.00	7.43	7.26	6.58	6.67	8.56	8.31	8.32	8.09
[Fe ₄ S ₄ (SCH ₂ CO ₂ Et) ₄] ²⁻	4*Fe ^{2.5+}	0.0	9.00	7.20	7.08	6.13	6.18	n/a ^c	8.21	n/a ^c	7.97
[Fe ₄ (NO) ₄ (<i>μ</i> ₃ -S) ₄]	4*{FeNO} ⁷	0.0	6.00	1.70	0.60	0.23	0.32	7.12	5.15	5.58	4.07
[Fe ₄ (NO) ₄ (<i>μ</i> ₃ -S) ₄] ⁻	2*{FeNO} ^{7.5} 2*{FeNO} ⁷	0.75	5.75	3.13	1.82	0.85	1.24	7.46	6.50	5.88	5.44
MAD ($\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$)				2.08	2.51	3.25	3.12	0.75	0.52	0.69	0.92

^a See also the Supporting Information, Table S6, for the B3LYP and B3LYP* values using ORCA in combination with a CP(PPP) basis set on Fe and a TZVP basis set on the remaining atoms. ^b For the mononuclear Fe complexes, the expectation value of S^2 is for the unrestricted, rather than broken symmetry, state. WC refers to weak-coupling limit. ^c The OLYP/STO-TZP COSMO geometry is used instead. See the Computational Methods section for details.

occupied orbitals can be written explicitly as

$$\langle \hat{S}^2 \rangle_{\text{BS}} = \left(\frac{N^\alpha - N^\beta}{2} \right) \left(\frac{N^\alpha - N^\beta}{2} + 1 \right) + N^\beta - \sum_i n_{iD} |\tilde{S}_{iD}|^2 - \sum_i n_{iW} |\tilde{S}_{iW}|^2 \quad (6)$$

where $|\tilde{S}_{iD}|^2$ are the corresponding orbital overlaps squared close to 1 (near double occupancy) and $|\tilde{S}_{iW}|^2$ are the corresponding orbital overlaps squared near 0 (weakly coupled). Substituting eq 5 into 6 gives

$$\langle \hat{S}^2 \rangle_{\text{BS}} = \left(\frac{N_W^\alpha - N_W^\beta}{2} \right) \left(\frac{N_W^\alpha - N_W^\beta}{2} + 1 \right) + N_W^\beta + [N_D - \sum_i n_{iD} (\tilde{S}_{iD})^2] - \sum_i n_{iW} (\tilde{S}_{iW})^2 \quad (7)$$

In the weak-coupling limit, the contribution from doubly occupied orbitals in eq 7 is neglected since the third term (in the square brackets) is small and positive and the fourth term is small and negative. In an analogous way, for a system comprised of spin subsystems A and B that couple antiferromagnetically, we can assume that the sites couple weakly to give the spin expectation value as

$$\langle S^2 \rangle_{\text{BS}}^{\text{WC}} = S_{\text{max}}(S_{\text{max}} + 1) - 4S_A S_B = (S_A - S_B)(S_A + S_B + 1) + 2S_B \quad (8)$$

where $S_{\text{max}} = S_A + S_B$, $S_A = N_W^\alpha/2$, and $S_B = N_W^\beta/2$. It is easily seen that eq 8 corresponds precisely to the first two terms in eq 7. The weak coupling limit approximation and its utility for understanding spin states was described earlier.^{82,83}

As can be seen in Table 6, the calculated BS values for different exchange-correlation potentials ($\langle S^2 \rangle_{\text{BS}}^{\text{XC}}$) do not always correspond to the pure spin-state values ($\langle S^2 \rangle_{\text{pure}}$) or to the BS values in the weak-coupling limit ($\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$). The three Fe–S mononuclear complexes exhibit good agreement between the calculated and pure spin expectation values. Less agreement is observed, however, between the calculated versus pure, $\langle S^2 \rangle_{\text{pure}}$, or weak-coupling limit spin expectation values, $\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$, for the four Fe–NO mononuclear complexes. With respect to $\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$, these differences can be attributed to our assumption that the Fe–NO unit is bonded covalently, but that it is comprised of high-spin Fe³⁺ ($S = 5/2$) coupled antiferromagnetically with NO⁻ ($S = 1$) to give a site spin for Fe–NO of 3/2. Moreover, in principle, several values of $\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$ can be constructed, depending on how the Fe–NO unit is described. One notable aspect of our calculated S^2 values is the different behaviors of the GGA versus hybrid exchange-correlation functionals. Although all levels of theory give values below $\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$ for the mononuclear Fe complexes, B3LYP and B3LYP* give values closer to the weak-coupling limit. Similarly, B3LYP and B3LYP* predict higher values of S^2 for the di- and tetranuclear Fe complexes. Structurally, this can be understood in terms of the longer bond lengths obtained with these functionals versus the GGA functionals (see, e.g., Tables 2–4). The calculated S^2 values for the polynuclear Fe–NO

complexes demonstrate remarkable variability versus $\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$, which highlights their challenging electronic structure. In addition to the strong coupling that we assume in the Fe–NO units, the coupling between the Fe–NO units in the polynuclear complexes is often strong, as observed elsewhere,^{11,12} which leads to a breakdown in the weak-coupling limit approximation.

Given the rather poor geometries obtained in the present study with B3LYP and B3LYP*, it is possible that spin-projection techniques would give structures in closer agreement to experimental results, as observed elsewhere.^{78,79} However, since we find that geometries obtained with OLYP/STO-TZP are in very good agreement with experimental structures (see, e.g., Table 2), we examined the effect of calculating B3LYP and B3LYP* linear

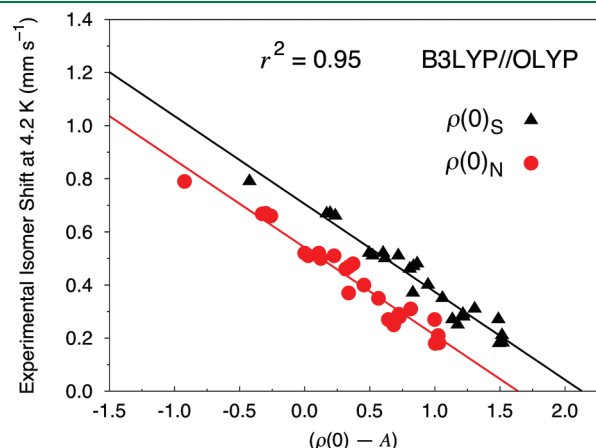


Figure 5. Isomer shift fit based on eight Fe–NO and 12 Fe–S complexes (24 sites) calculated at the B3LYP/STO-TZP//OLYP/STO-TZP COSMO level of theory. The Fe electron density is calculated directly at the nucleus ($\rho(0)_{\text{N}}$, circles) and on a small sphere around the center of the Fe nucleus ($\rho(0)_{\text{S}}$, triangles).

correlation fits and calculated Mössbauer properties using OLYP geometries (denoted B3LYP//OLYP and B3LYP*//OLYP, respectively). It is quite possible that similar results would be obtained with Gaussian-type orbitals, subject to OLYP with a GTO basis set providing accurate geometries.

Figure 5 shows that the data obtained with B3LYP//OLYP are quite an improvement over the calibration fits acquired using B3LYP geometries (Figure S1, Supporting Information). That is, fits derived using the entire test set improve the r^2 from 0.64 to 0.95, and the MAE decreases from 0.086 to 0.031 mm s^{-1} . Universal fit parameters for B3LYP//OLYP are $\alpha = -0.331$, $C = 0.705$, and $A = 11823$ when the Fe electron density is calculated on a small sphere surrounding the nucleus ($\rho(0)_{\text{S}}$) and $\alpha = -0.330$, $C = 0.541$, and $A = 11880$ when the Fe electron density is calculated at the nucleus ($\rho(0)_{\text{N}}$, Table 4). The largest outlier is for $[\text{Fe}(\text{SET})_4]^-$, which possesses a MAE of 0.07 mm s^{-1} . It is interesting to note that the fits derived from B3LYP//OLYP calculations are of superior quality to those obtained with OLYP alone. Indeed, the r^2 and MAE for B3LYP//OLYP (0.95 and 0.031 mm s^{-1} , respectively) reflect a slightly improved performance over OLYP (0.92 and 0.037 mm s^{-1}).

Similar findings occur for fits derived from B3LYP*//OLYP calculations compared with those using B3LYP* geometries (Figure 3 and Figure S16, Supporting Information). In this case, the r^2 improves from 0.86 to 0.95 and the MAE decreases from 0.053 mm s^{-1} to 0.032 mm s^{-1} when OLYP geometries are used instead of those calculated with B3LYP*. As found for B3LYP//OLYP, the largest outlier is for $[\text{Fe}(\text{SET})_4]^-$, which in this case possesses a MAE of 0.08 mm s^{-1} . Table 7 and Figure S16 of the Supporting Information provide the universal fit parameters and plot for this level of theory.

Grouping the test set into non-nitrosyl (Fe–S) and iron–nitrosyl (Fe–NO) complexes gives results similar to what we have seen previously (Table 5). Using OLYP geometries instead of B3LYP improves the r^2 from 0.80 to 0.92 and decreases the

Table 7. Universal Fit Parameters for the Calculation of ^{57}Fe Isomer Shifts^a

functional	method to calculate $\rho(0)^b$	α	C	A	r^2	MAE (mm s^{-1})
OLYP	S	-0.324	0.594	11820	0.92	0.037
	N	-0.323	0.428	11877		
OPBE	S	-0.287	0.594	11820	0.89	0.041
	N	-0.286	0.447	11877		
PW91	S	-0.334	0.703	11827	0.89	0.042
	N	-0.332	0.549	11884		
BP86	S	-0.354	0.869	11832	0.91	0.040
	N	-0.353	0.718	11889		
B3LYP	S	-0.340	0.633	11823	0.64	0.086
	N	-0.338	0.465	11880		
B3LYP//OLYP	S	-0.331	0.705	11823	0.95	0.031
	N	-0.330	0.541	11880		
B3LYP*	S	-0.403	0.609	11820	0.86	0.053
	N	-0.401	0.804	11876		
B3LYP*//OLYP	S	-0.337	0.633	11820	0.95	0.032
	N	-0.335	0.796	11876		
B3LYP ^{sc}	N	-0.374	0.489	11815	0.77	0.064
B3LYP ^c	N	-0.347	0.623	11816	0.69	0.079

^a Using COSMO in combination with a STO-TZP basis set unless otherwise indicated. ^b The S serves to indicate that the Fe electron density has been calculated on a small sphere around the center of the Fe nucleus, while the N indicates that the density has been calculated at the Fe nucleus. ^c Using COSMO in combination with the CP(PPP) basis set for Fe and the TZVP basis set for the remaining elements.

Table 8. Fit Parameters for the Calculation of ^{57}Fe Isomer Shifts^a

functional	fit type	method to calculate $\rho(0)^b$	α	C	A	r^2	MAE (mm s^{-1})
OLYP	Fe–S	S	−0.465	0.652	11820	0.88	0.031
		N	−0.463	0.413	11877		
	Fe–NO	S	−0.295	0.572	11820		
		N	−0.294	0.421	11877		
OPBE	Fe–S	S	−0.461	0.670	11820	0.83	0.037
		N	−0.459	0.433	11877		
	Fe–NO	S	−0.263	0.577	11820		
		N	−0.262	0.442	11877		
PW91	Fe–S	S	−0.485	0.787	11827	0.84	0.041
		N	−0.482	0.564	11884		
	Fe–NO	S	−0.329	0.722	11827		
		N	−0.328	0.570	11884		
BP86	Fe–S	S	−0.495	1.012	11832	0.86	0.038
		N	−0.492	0.800	11889		
	Fe–NO	S	−0.343	0.871	11832		
		N	−0.342	0.725	11889		
B3LYP	Fe–S	S	−0.279	0.651	11823	0.80	0.045
		N	−0.278	0.513	11880		
	Fe–NO	S	−0.503	0.642	11823		
		N	−0.501	0.393	11880		
B3LYP//OLYP	Fe–S	S	−0.372	0.737	11823	0.92	0.028
		N	−0.370	0.553	11880		
	Fe–NO	S	−0.314	0.685	11823		
		N	−0.313	0.529	11880		
B3LYP*	Fe–S	S	−0.346	0.620	11820	0.90	0.034
		N	−0.344	0.787	11876		
	Fe–NO	S	−0.441	0.577	11820		
		N	−0.439	0.791	11876		
B3LYP*//OLYP	Fe–S	S	−0.407	0.668	11820	0.92	0.027
		N	−0.405	0.864	11876		
	Fe–NO	S	−0.316	0.618	11820		
		N	−0.315	0.771	11876		
B3LYP ^c	Fe–S	N	−0.285	0.641	11816	0.82	0.044
	Fe–NO	N	−0.476	0.614	11816	0.93	0.048
B3LYP ^{nc}	Fe–S	N	−0.326	0.529	11815	0.84	0.037
	Fe–NO	N	−0.435	0.428	11815	0.93	0.045

^a Using COSMO in combination with a STO-TZP basis set unless otherwise indicated. ^b The S serves to indicate that the Fe electron density has been calculated on a small sphere around the center of the Fe nucleus, while the N indicates the density has been calculated at the Fe nucleus. ^c Using COSMO in combination with the CP(PPP) basis set for Fe and the TZVP basis set for the remaining elements.

MAE from 0.045 mm s^{-1} to 0.028 mm s^{-1} for the Fe–S complexes. For the Fe–NO complexes, the r^2 improves from 0.91 to 0.97 and the MAE decreases from 0.045 mm s^{-1} to 0.030 mm s^{-1} . The largest IS outliers for the Fe–S and Fe–NO complexes, each with MAEs of 0.06 mm s^{-1} , correspond to the Fe–SPh site in $[\text{Fe}_4\text{S}_4(\text{SPh})_2\text{Cl}_2]^{2-}$ and the $\{\text{FeNO}\}^{7,5}$ site in $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$ and $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$.

For the B3LYP*//OLYP calculations of the Fe–S complexes, the r^2 improves from 0.90 to 0.92 and the MAE decreases from 0.034 mm s^{-1} to 0.027 mm s^{-1} . The largest outlier in this group is $[\text{Fe}_2\text{S}_2(\text{OPh-}p\text{-CH}_3)_4]^{2-}$, which possesses a MAE of 0.07 mm s^{-1} . For the Fe–NO complexes, the r^2 improves slightly from 0.96 to 0.97 and the MAE decreases from 0.033 mm s^{-1} to 0.028 mm s^{-1} with B3LYP*//OLYP. The largest outlier in this case is for the $\{\text{FeNO}\}^7$ and $\{\text{FeNO}\}^{7,5}$ sites in $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$ (MAEs of 0.05 mm s^{-1}).

In Tables 7 and 8, we summarize the linear regression fitting parameters obtained in this study. Table 7 provides universal fit parameters derived from all 20 structures within our test set and Table S3 of the Supporting Information gives the calculated isomer shifts for these complexes using these universal fitting parameters. Overall, the OLYP, OPBE, PW91, and BP86 functionals perform very well, with the hybrid method B3LYP* performing slightly worse, and B3LYP performing worse still. However, the use of OLYP geometries to calculate B3LYP and B3LYP* isomer shift parameters significantly improves the correlation analysis. In addition, the performance of B3LYP and B3LYP* increases substantially when the non-nitrosyl (Fe–S) and iron–nitrosyl (Fe–NO) compounds of our test set are grouped separately (Table 8). Indeed, when separate calibration lines are generated for the non-nitrosyl (Fe–S) and iron–nitrosyl (Fe–NO) complexes, very good to excellent performance is observed for all levels of theory.

Table 9. Fit Parameters for the Calculation of ^{57}Fe Quadrupole Splittings^a

functional	fit type	α	C	r^2
OLYP	universal	1.095	0.138	0.82
	Fe–S	1.278	–0.030	0.89
	Fe–NO	0.617	0.580	0.80
	Fe–NO ^c	0.646	0.612	0.92
OPBE	universal	1.088	0.161	0.83
	Fe–S	1.299	–0.034	0.91
	Fe–NO	0.613	0.593	0.88
	Fe–NO ^c	0.597	0.645	0.91
PW91	universal	1.055	0.227	0.94
	Fe–S	1.173	0.084	0.98
	Fe–NO	0.663	0.603	0.89
	Fe–NO ^c	0.651	0.641	0.92
BP86	universal	1.051	0.223	0.94
	Fe–S	1.169	0.078	0.99
	Fe–NO	0.664	0.599	0.91
	Fe–NO ^c	0.646	0.646	0.93
B3LYP	universal	0.808	0.137	0.89
	Fe–S	0.825	0.074	0.91
	Fe–NO	0.712	0.310	0.74
	Fe–NO ^c	0.823	0.207	0.90
B3LYP//OLYP	universal	0.849	0.131	0.91
	Fe–S	0.890	0.076	0.95
	Fe–NO	0.591	0.450	0.66
	Fe–NO ^c	0.708	0.387	0.91
B3LYP*	universal	0.875	0.115	0.93
	Fe–S	0.909	0.020	0.96
	Fe–NO	0.705	0.377	0.82
	Fe–NO ^c	0.814	0.262	0.95
B3LYP*//OLYP	universal	0.874	0.163	0.92
	Fe–S	0.931	0.078	0.97
	Fe–NO	0.577	0.519	0.69
	Fe–NO ^c	0.675	0.484	0.92
B3LYP ^b	universal	0.846	0.082	0.82
	Fe–S	0.883	–0.031	0.86
	Fe–NO	0.653	0.397	0.57
	Fe–NO ^c	0.861	0.259	0.92
B3LYP ^b	universal	0.917	0.027	0.82
	Fe–S	0.977	–0.122	0.86
	Fe–NO	0.661	0.413	0.65
	Fe–NO ^c	0.814	0.341	0.97

^aUsing COSMO in combination with a STO-TZP basis set unless otherwise indicated. ^bUsing COSMO in combination with the CP-(PPP) basis set for Fe and the TZVP basis set for the remaining elements. ^cExcluding $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$.

This behavior is reminiscent of earlier observations that different Fe oxidation states yield distinct IS calibration lines.^{11,24,26} In the present case, the separate grouping appears to be very important when B3LYP* and B3LYP geometries are used. A defining feature of the separate calibration analysis for all levels of theory is the excellent linear correlation of experimental ISs and the calculated Fe nuclear density for the Fe–NO complexes ($r^2 = 0.91–0.98$), with slightly worse correlation for the non-nitrosyl Fe–S compounds ($r^2 = 0.80–0.92$).

Good correspondence between the fit parameters determined here and those determined previously by our group has been found. Using an identical test set to the one used here, universal fit parameters for the OLYP functional (with COSMO) were determined to be $\alpha = -0.315$ and $C = 0.432$, with an r^2 of 0.915 and a MAE of 0.039 mm s^{-1} when the Fe electron density is calculated at the nucleus.¹¹ Fit parameters for Fe–NO complexes only were found to be $\alpha = -0.290$ and $C = 0.426$, with an r^2 of 0.979 and a MAE of 0.029 mm s^{-1} . Both sets of these universal and Fe–NO fit parameters compare extremely well with those determined herein with the OLYP functional (Tables 7 and 8). In addition, earlier Mössbauer isomer fitting procedures on a test set comprised of 19 $\text{Fe}^{2.5+,3+,3.5+,4+}$ complexes (with 30 Fe sites) yield fit parameters in good agreement with those determined here.^{23,24} For example, fits derived from OLYP calculations give parameters $\alpha = -0.307$ and $C = 0.385$ with an r of -0.93 , while fits derived from OPBE calculations give $\alpha = -0.312$ and $C = 0.373$ ($r = -0.94$) and fits derived from PW91 calculations give $\alpha = -0.393$ and $C = 0.435$ ($r = -0.93$). All of these earlier calculations determined the Fe electron density directly at the nucleus.

c. Quadrupole Splittings. As noted in the Introduction, quadrupole splitting (QS) arises from the interaction between the electric quadrupole moment of the ^{57}Fe nucleus and the electric field gradient at its nucleus. Because very few of the signs of the QSs within our test set have been determined experimentally, we quote only their magnitude. Tables S4 and S5 in the Supporting Information show our calculated values for the QS and η values. When one experimental QS value is quoted for polynuclear complexes, our calculated values were averaged over equivalent Fe centers. Table S4 shows B3LYP*/STO-TZP to yield the lowest mean absolute error (MAE) relative to experimental results (0.19 mm s^{-1}) for the calculation of QS parameters, followed by B3LYP*//OLYP (0.20 mm s^{-1}) and B3LYP//OLYP (0.21 mm s^{-1}) with the STO-TZP basis set; B3LYP/STO-TZP (0.26 mm s^{-1}), B3LYP, and B3LYP* with the CP-(PPP)-TZVP basis set (0.27 mm s^{-1} each); and OPBE (0.28 mm s^{-1}), OLYP (0.29 mm s^{-1}), BP86 (0.30 mm s^{-1}), and PW91 (0.30 mm s^{-1}) with STO-TZP basis set. These calculated values are prior to any linear fitting. See Table 1 for experimental $|\text{QS}|$ values. It is noteworthy that the hybrid DFT functionals tend to perform better at calculating QS parameters than the pure functionals examined here.

Next, we examined linear correlations between calculated and observed QS absolute values based on the equation:

$$|\text{QS}_{(\text{exp})}| = \alpha|\text{QS}_{(\text{calc})}| + C \quad (3)$$

Universal fit parameters were obtained by linear regression of our calculated QS parameters of 24 distinct Fe sites in the 20 complexes of our test set (Table 9). As we have done in our IS calibration analysis, we also evaluated fit parameters for test sets including non-nitrosyl (Fe–S) and iron–nitrosyl (Fe–NO) complexes separately. Using r^2 as a measure of the quality of the calibration fit, most of the functionals perform very well, especially BP86. For the universal fitting procedure, OLYP/STO-TZP and B3LYP and B3LYP* with the CP-(PPP)-TZVP basis set give fits with the worst r^2 values (0.82). Interestingly, B3LYP//OLYP and B3LYP*//OLYP yield calibration fits substantially improved over the fit obtained from OLYP (with r^2 values of 0.91 and 0.92, respectively).

The worst fits are obtained from the B3LYP and B3LYP* calculations with the CP-(PPP)-TZVP basis set for the Fe–NO

complexes, with r^2 values of 0.57 and 0.65, respectively (Table 9). Though not as severe, the calibration fits for the Fe–NO complexes derived from our B3LYP//OLYP and B3LYP*//OLYP calculations give correlation coefficients of 0.66 and 0.69, respectively. The rather poor performance of B3LYP and B3LYP* is also observed with the STO-TZP basis set (r^2 values of 0.74 and 0.82, respectively). The largest outlier in these calculations is $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$, and its removal gives rather impressive results. For instance, r^2 improves from 0.57 to 0.92 for B3LYP/CP(PPP)-TZVP, from 0.65 to 0.97 for B3LYP*/CP(PPP)-TZVP, from 0.66 to 0.91 for B3LYP//OLYP, and from 0.69 to 0.92 for B3LYP*//OLYP (Table 9). Indeed, the correlation coefficient is improved for every level of theory when $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$ is excluded.

Initially, we thought that the source of the rather poor results for $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$ may be a consequence of poorly calculated geometries. For example, B3LYP and B3LYP* tend to overestimate the average Fe–Fe bond distances by approximately 0.2 Å and the average Fe–N bond distances by approximately 0.1 Å (Table 2). Such deviations, however, are not observed in the geometries optimized with OLYP, OPBE, PW91, or BP86, yet their calibration fits are also improved upon the exclusion of $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$. Therefore, while the geometry clearly plays a role in determining accurate QSs, the inherent nature of the functional used to calculate this property is also of great importance.

CONCLUSIONS

Density functional theory calculations have been performed on a test set of 20 compounds including nonheme nitrosyls (Fe–NO) and non-nitrosyl (Fe–S) complexes to calibrate Mössbauer isomer shift (IS) and quadrupole splitting (QS) parameters for general use. Fits to determine accurate ISs were obtained by calculating the Fe electron density on a small sphere surrounding its nucleus and directly at the nucleus using the OLYP, OPBE, PW91, BP86, B3LYP, and B3LYP* functionals. In addition, the latter two functionals were used in conjunction with Slater-type orbitals and Gaussian-type orbitals. We find that the methods of calculating the Fe electron density directly at or on a small sphere surrounding the nucleus yield (near) identical slopes from the linear regression analyses but are shifted with respect to $\rho(0)$ along the x axis.

Universal fit parameters for the calculation of ^{57}Fe isomer shifts obtained with the OLYP, OPBE, PW91, and BP86 functionals are found to be of very good quality, while those obtained with B3LYP and B3LYP* do not produce the same level of accuracy. We attribute the inferior performance of B3LYP and B3LYP* to poorly calculated geometries with these functionals. This observation has potentially large consequences since B3LYP, in particular, is used widely. Moreover, we note that the largest discrepancies have been observed for the di- and tetranuclear Fe complexes within our training set. Given these findings, it is uncertain whether previously determined calibration fits for the prediction of Mössbauer isomer shifts that have included mostly mononuclear Fe complexes are appropriate for the prediction of isomer shifts in polynuclear Fe complexes. The inclusion of effects due to dispersion via the B3LYP-D functional is found to slightly improve the geometries for some mono- and dinuclear Fe complexes, but this functional performs worse than B3LYP for the tetranuclear Fe complexes.

An examination of spin expectation values reveals that B3LYP and B3LYP* approach the weak-coupling limit more closely than

the GGA exchange-correlation functionals, which can be understood to arise from the longer bond lengths obtained with these hybrid functionals. The Fe–NO complexes demonstrate large variability in the calculated S^2 values relative to $\langle S^2 \rangle_{\text{BS}}^{\text{WC}}$, which is probably a consequence of assuming that the Fe–NO units are bound covalently though coupled antiferromagnetically. Further, for the polynuclear complexes, the strong coupling between the Fe–NO units can lead to a breakdown in the weak-coupling limit approximation, especially for the GGA functionals.

Interestingly, we find that the use of OLYP geometries for the B3LYP and B3LYP* calculations with an STO-TZP basis set (B3LYP//OLYP and B3LYP*//OLYP, respectively) significantly improves the quality of the calibration fits, providing further evidence that the structures obtained with B3LYP and B3LYP* for the complexes within the test set are not reliable. Indeed, we find B3LYP and B3LYP* to frequently overestimate the Fe–Fe and Fe–S distances in the structures of our test set (Tables 2 and 3).

In addition, we have also determined isomer shift calibration fits for test sets comprised of only Fe–NO or Fe–S complexes. Every functional examined gives fits of very good quality, though those determined for the Fe–S complexes tend to demonstrate poorer correlations. As found with the universal fitting parameters, we find that the calibration fits for B3LYP//OLYP and B3LYP*//OLYP are superior to those determined using the B3LYP and B3LYP* geometries. These data suggest that the composition of a given test set can strongly influence the correlation between experimental ISs and calculated Fe electron densities. This appears to be especially true for B3LYP and B3LYP*.

Universal and separate Fe–NO and Fe–S fit parameters obtained to determine QSs are found to be of good to excellent quality for every density functional examined, especially if $[\text{Fe}_4(\text{NO})_4(\mu_3\text{-S})_4]^-$ is removed from the test set.

ASSOCIATED CONTENT

S Supporting Information. Geometries calculated at the OLYP, OPBE, PW91, BP86, B3LYP, and B3LYP* level; Mulliken spin populations (Tables S1 and S2), calculated isomer shifts (Table S3); calculated quadrupole splittings (Table S4); η values (Table S5); and various isomer shift fits (Figures S1–S18). This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: lou@scripps.edu.

ACKNOWLEDGMENT

We gratefully acknowledge financial support through NIH grant GM039914 (to L.N.), the Research Council of Norway and a YFF grant (to A.G.), and computer resources from the Scripps Research Institute. We also wish to thank Professor Frank Neese for providing us with the ORCA code and helpful ADF discussions with Dr. Erik van Lenthe of SCM. Finally, we wish to thank an anonymous reviewer for encouraging us to examine the performance of B3LYP-D to calculate geometries.

REFERENCES

(1) Schünemann, V.; Paulsen, H. *Mössbauer Spectroscopy. In Applications of Physical Methods to Inorganic and Bioinorganic Chemistry*;

- Scott, R. A.; Lukehart, C. M., Eds.; John Wiley & Sons Ltd: Chichester, U. K., 2007; pp 243–69.
- (2) Krebs, C.; Price, J. C.; Baldwin, J.; Saleh, L.; Green, M. T.; Bollinger, J. M. *Inorg. Chem.* **2005**, *44*, 742–57.
- (3) Neese, F. *Inorg. Chim. Acta* **2002**, *337*, 181–92.
- (4) Neese, F. *Coord. Chem. Rev.* **2009**, *253*, 526–63.
- (5) Siegbahn, P. E. M. *J. Comput. Chem.* **2001**, *22*, 1634–45.
- (6) Noodleman, L.; Lovell, T.; Han, W. G.; Li, J.; Himmo, F. *Chem. Rev.* **2004**, *104*, 459–508.
- (7) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 949–1017.
- (8) Enemark, J. H.; Feltham, R. D. *Coord. Chem. Rev.* **1974**, *13*, 339–406.
- (9) Ghosh, A. *Acc. Chem. Res.* **2005**, *38*, 943–54.
- (10) McCleverty, J. A. *Chem. Rev.* **2004**, *104*, 403–18.
- (11) Hopmann, K. H.; Ghosh, A.; Noodleman, L. *Inorg. Chem.* **2009**, *48*, 9155–65.
- (12) Hopmann, K. H.; Noodleman, L.; Ghosh, A. *Chem.—Eur. J.* **2010**, *16*, 10397–408.
- (13) Zhang, Y.; Oldfield, E. *J. Phys. Chem. A* **2003**, *107*, 4147–50.
- (14) Conradie, J.; Quarless, D. A.; Hsu, H. F.; Harrop, T. C.; Lippard, S. J.; Koch, S. A.; Ghosh, A. *J. Am. Chem. Soc.* **2007**, *129*, 10446–56.
- (15) Conradie, J.; Ghosh, A. *J. Phys. Chem. B* **2007**, *111*, 12621–24.
- (16) Hopmann, K. H.; Conradie, J.; Ghosh, A. *J. Phys. Chem. B* **2009**, *113*, 10540–47.
- (17) Nemykin, V. N.; Kobayashi, N.; Chernii, V. Y.; Belsky, V. K. *Eur. J. Inorg. Chem.* **2001**, 733–43.
- (18) Zhang, Y.; Mao, J. H.; Oldfield, E. *J. Am. Chem. Soc.* **2002**, *124*, 7829–39.
- (19) Nemykin, V. N.; Hadt, R. G. *Inorg. Chem.* **2006**, *45*, 8297–307.
- (20) Zhang, Y.; Mao, J. H.; Godbout, N.; Oldfield, E. *J. Am. Chem. Soc.* **2002**, *124*, 13921–30.
- (21) Liu, T. Q.; Lovell, T.; Han, W. G.; Noodleman, L. *Inorg. Chem.* **2003**, *42*, 5244–51.
- (22) Sinnecker, S.; Slep, L. D.; Bill, E.; Neese, F. *Inorg. Chem.* **2005**, *44*, 2245–54.
- (23) Han, W. G.; Liu, T. Q.; Lovell, T.; Noodleman, L. *J. Comput. Chem.* **2006**, *27*, 1292–306.
- (24) Han, W. G.; Noodleman, L. *Inorg. Chim. Acta* **2008**, *361*, 973–86.
- (25) Römelt, M.; Ye, S. F.; Neese, F. *Inorg. Chem.* **2009**, *48*, 784–85.
- (26) Bochevarov, A. D.; Friesner, R. A.; Lippard, S. J. *J. Chem. Theory Comput.* **2010**, *6*, 3735–49.
- (27) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–11.
- (28) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–87.
- (29) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524–29.
- (30) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–24.
- (31) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–12.
- (32) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–89.
- (33) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–68.
- (34) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–96.
- (35) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–52.
- (36) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–27.
- (37) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48–55.
- (38) ADF2010.02; SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands. See <http://www.scm.com> (accessed Aug 2011).
- (39) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–67.
- (40) Neese, F. ORCA, version 2.8.0; University of Bonn: Bonn, Germany, 2010. <http://www.thch.uni-bonn.de/tc/orca> (accessed Aug 2011).
- (41) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–77.
- (42) Ahlrichs, R. and co-workers. Unpublished.
- (43) Wachters, A. J. *J. Chem. Phys.* **1970**, *52*, 1033–36.
- (44) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.
- (45) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737–43.
- (46) Liu, T.; Noodleman, L.; Case, D. A. *hyper2003*; The Scripps Research Institute: La Jolla, CA.
- (47) Martinez-Pinedo, G.; Schwerdtfeger, P.; Caurier, E.; Langanke, K.; Nazarewicz, W.; Sohnel, T. *Phys. Rev. Lett.* **2001**, *87*, 062701.
- (48) Zhang, Y.; Pavlosky, M. A.; Brown, C. A.; Westre, T. E.; Hedman, B.; Hodgson, K. O.; Solomon, E. I. *J. Am. Chem. Soc.* **1992**, *114*, 9189–91.
- (49) Conradie, J.; Ghosh, A. *Inorg. Chem.* **2011**, *50*, 4223–25.
- (50) Conradie, J.; Hopmann, K. H.; Ghosh, A. *J. Phys. Chem. B* **2010**, *114*, 8517–24.
- (51) Ghosh, A.; Conradie, J.; Hopmann, K. H. Electronic Structure Calculations: Transition Metal–NO Complexes. In *Computational Inorganic and Bioinorganic Chemistry*; Solomon, E. I., King, R. B., Scott, R. A., Eds.; John Wiley & Sons, Ltd.: Chichester, U. K., 2009; pp 389–410.
- (52) Maelia, L. E.; Millar, M.; Koch, S. A. *Inorg. Chem.* **1992**, *31*, 4594–600.
- (53) Macdonnell, F. M.; Ruhlandtsenge, K.; Ellison, J. J.; Holm, R. H.; Power, P. P. *Inorg. Chem.* **1995**, *34*, 1815–22.
- (54) Coucouvanis, D.; Swenson, D.; Baenziger, N. C.; Murphy, C.; Holah, D. G.; Sfarnas, N.; Simopoulos, A.; Kostikas, A. *J. Am. Chem. Soc.* **1981**, *103*, 3350–62.
- (55) Wanat, A.; Schnepf, T.; Stochel, G.; van Eldik, R.; Bill, E.; Wieghardt, K. *Inorg. Chem.* **2002**, *41*, 4–10.
- (56) Johnson, C. E.; Rickards, R.; Hill, H. A. O. *J. Chem. Phys.* **1969**, *50*, 2594–97.
- (57) Butcher, R. J.; Sinn, E. *Inorg. Chem.* **1980**, *19*, 3622–26.
- (58) Harrop, T. C.; Tonzetich, Z. J.; Reisner, E.; Lippard, S. J. *J. Am. Chem. Soc.* **2008**, *130*, 15602–10.
- (59) Sanina, N. A.; Rakova, O. A.; Aldoshin, S. M.; Shilov, G. V.; Shulga, Y. M.; Kulikov, A. V.; Ovanesyan, N. S. *Mendeleev Commun.* **2004**, 7–8.
- (60) Sanina, N. A.; Aldoshin, S. M. *Russ. Chem. Bull.* **2004**, *53*, 2428–48.
- (61) Mayerle, J. J.; Denmark, S. E.; Depamphilis, B. V.; Ibers, J. A.; Holm, R. H. *J. Am. Chem. Soc.* **1975**, *97*, 1032–45.
- (62) Gillum, W. O.; Frankel, R. B.; Foner, S.; Holm, R. H. *Inorg. Chem.* **1976**, *15*, 1095–100.
- (63) Salifoglou, A.; Simopoulos, A.; Kostikas, A.; Dunham, R. W.; Kanatzidis, M. G.; Coucouvanis, D. *Inorg. Chem.* **1988**, *27*, 3394–406.
- (64) Feig, A. L.; Bautista, M. T.; Lippard, S. J. *Inorg. Chem.* **1996**, *35*, 6892–98.
- (65) Davies, S. C.; Evans, D. J.; Hughes, D. L.; Konkol, M.; Richards, R. L.; Sanders, J. R.; Sobota, P. *J. Chem. Soc., Dalton Trans.* **2002**, 2473–82.
- (66) Kanatzidis, M. G.; Baenziger, N. C.; Coucouvanis, D.; Simopoulos, A.; Kostikas, A. *J. Am. Chem. Soc.* **1984**, *106*, 4500–11.
- (67) Excoffon, P.; Laugier, J.; Lamotte, B. *Inorg. Chem.* **1991**, *30*, 3075–81.
- (68) Cleland, W. E.; Holtman, D. A.; Sabat, M.; Ibers, J. A.; Defotis, G. C.; Averill, B. A. *J. Am. Chem. Soc.* **1983**, *105*, 6021–31.
- (69) Segal, B. M.; Hoveyda, H. R.; Holm, R. H. *Inorg. Chem.* **1998**, *37*, 3440–43.
- (70) Silver, J.; Fern, G. R.; Miller, J. R.; McCammon, C. A.; Evans, D. J.; Leigh, G. J. *Inorg. Chem.* **1999**, *38*, 4256–61.
- (71) Sedney, D.; Reiff, W. M. *Inorg. Chim. Acta* **1979**, *34*, 231–36.
- (72) Baird, P.; Bandy, J. A.; Green, M. L. H.; Hamnett, A.; Marseglia, E.; Obertelli, D. S.; Prout, K.; Qin, J. G. *J. Chem. Soc., Dalton Trans.* **1991**, 2377–93.

- (73) Chu, C. T. W.; Lo, F. Y. K.; Dahl, L. F. *J. Am. Chem. Soc.* **1982**, *104*, 3409–22.
- (74) Fee, J. A.; Findling, K. L.; Yoshida, T.; Hille, R.; Tarr, G. E.; Hearshen, D. O.; Dunham, W. R.; Day, E. P.; Kent, T. A.; Munck, E. *Biol. Chem.* **1984**, *259*, 124–33.
- (75) Radon, M.; Broclawik, E.; Pierloot, K. *J. Phys. Chem. B* **2010**, *114*, 1518–28.
- (76) Radon, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824–32.
- (77) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–99.
- (78) Li, J.; Noodleman, L. Electronic structure calculations: Density functional methods for spin polarization, charge transfer, and solvent effects in transition metal complexes. In *Spectroscopic Methods in Bioinorganic Chemistry*; Solomon, E. I., Hodgson, K. O., Eds.; American Chemical Society: Washington DC, 1998; pp 179–95.
- (79) Edgecombe, K. E.; Becke, A. D. *Chem. Phys. Lett.* **1995**, *244*, 427–32.
- (80) Szabo, A.; Ostlund, N. S. *Modern Theoretical Chemistry: Introduction to Advanced Electronic Structure Theory*; McGraw-Hill, Inc.: New York, 1989.
- (81) Neese, F. *J. Phys. Chem. Solids* **2004**, *65*, 781–85.
- (82) Mouesca, J. M.; Chen, J. L.; Noodleman, L.; Bashford, D.; Case, D. A. *J. Am. Chem. Soc.* **1994**, *116*, 11898–914.
- (83) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J. M. *Coord. Chem. Rev.* **1995**, *144*, 199–244.

Magnetic Properties of Ni²⁺(aq) from First Principles

Jiří Mareš,^{*,†} Helmi Liimatainen,[‡] Teemu O. Pennanen,[‡] and Juha Vaara[†][†]NMR Research Group, Department of Physics, University of Oulu, P.O. Box 3000, FIN-90014, Oulu, Finland[‡]Laboratory of Physical Chemistry, Department of Chemistry, University of Helsinki, P.O. Box 55 (A. I. Virtasen aukio 1), FIN-00014, Helsinki, Finland Supporting Information

ABSTRACT: The aqueous solution of the Ni²⁺ ion was investigated using a first principles molecular dynamics (FPMD) simulation based on periodic density-functional theory (DFT) calculations. Statistical averages of the magnetic properties corresponding to the triplet spin state of the ion, the hyperfine coupling, *g* and zero-field splitting tensors, as well as the resulting paramagnetic nuclear magnetic resonance (pNMR) shielding terms were calculated using DFT from instantaneous simulation snapshots extracted from the FPMD trajectory. We report comprehensive tests of the reliability of systematically selected DFT functionals for the properties. The isotropic nuclear shielding of the ¹⁷O nuclei can be obtained with good predictive power. The accuracy of the calculated ¹H shieldings is limited by the fact that the spin-density on the proton sites is not reproduced reliably with the tested functionals, rendering the dominant Fermi contact isotropic shielding term less well-defined. On the other hand, the dominant spin-dipole term of the shielding anisotropy, which gives a practically vanishing isotropic contribution, can be obtained with good reliability for both the ¹H and ¹⁷O nuclei. The anisotropic shielding tensor can be thus utilized reliably in the calculation of Curie-type paramagnetic relaxation. We discuss the evolution of the pNMR properties through the first and second solvation shells of the ion, toward the bulk solvent. The magnetic properties of the dominant, six-coordinated solution are compared to those of the metastable, 5-fold coordinated intermediate occurring in the dissociative exchange process.

INTRODUCTION

A combined molecular dynamics (MD) simulation and molecular properties study provides insight into both static and time-dependent molecular properties not obtainable by static quantum-chemical models. By calculation of the molecular properties along the MD trajectory, both the finite-temperature effect as well as the influence of surrounding molecules (solvent) can be realistically incorporated.^{1–9} Furthermore, properties directly dependent on time such as the magnetic resonance relaxation can be modeled with a greatly decreased number of assumptions.¹⁰ In order to obtain results comparable with experiments, both the MD part as well as the property calculation steps require generally state-of-the-art methods, which usually results in time-consuming studies, thus limiting their number.

As an example of first-principles calculation of nuclear magnetic resonance properties in aqueous systems, Schmidt et al.¹⁰ investigated the quadrupolar NMR relaxation of heavy water using first-principles MD (FPMD) simulations, with the trajectory generated with the forces calculated “on the fly” using density-functional theory (DFT). There have only been a very limited number of studies concerning aqueous solutions of paramagnetic metal ions. Odelius et al.¹¹ calculated the average transient zero-field splitting (ZFS) along an empirical MD trajectory of Ni²⁺ in water, using a preparameterized ZFS hypersurface as a function of distortions from the idealized octahedral geometry. They used the unrestricted Hartree–Fock/restricted configuration interaction method to parameterize the ZFS hypersurface. Recently, Gd³⁺ in water was modeled by FPMD for the calculation of nuclear quadrupole¹² and hyperfine coupling constants (HFCs).¹³ Gd³⁺ in water has also been modeled using empirical

MD and combined with an electron spin resonance (ESR) relaxation study.¹⁴

The hydrated Ni²⁺ ion is a prototypical system for pNMR studies, which has been addressed in a wide range of investigations.^{11,15–20} Detailed understanding of paramagnetic relaxation enhancement (PRE) in such a model system would provide a link toward calculations of magnetic resonance imaging (MRI) contrast agents²¹ as well as applications in structural biology.²² To predict complicated phenomena such as NMR relaxation solely by calculation, it is crucial to be able to validate and recognize the impact of the many approximations that are necessary. In this work, we report an extensive computational study of the ESR and pNMR parameters of the Ni²⁺ ion in aqueous solution, using combined FPMD simulation and quantum-chemical calculations of simulation snapshots. A study of the structure and dynamics of the same system is reported separately.²³ The knowledge of the time evolution of these parameters is necessary for proceeding to dynamic NMR properties, which are experimentally manifested in NMR relaxation.²⁴ Comprehensive results of selected interactions such as transient ZFS and the *g* tensor for the triplet (*S* = 1) state system, and also a detailed breakdown of the theoretical²⁵ contributions to the average pNMR shielding of water nuclei, is presented both for static six- as well as five-coordinated Ni(H₂O)_{*n*}²⁺ (*n* = 5, 6) structures and the first and second solvation shells (FSS and SSS) of the dynamical FPMD trajectory. The two coordination cases represent the prevailing situations in the solution as well as the

Received: May 18, 2011

Published: August 04, 2011

Table 1. Basis-Set Dependence of the Orbital and Hyperfine ^1H and ^{17}O Nuclear Magnetic Shielding Constants (in ppm) in $\text{Ni}(\text{H}_2\text{O})_6^{2+}$ Calculated with Density-Functional Theory (PBE Functional)

basis set $\text{H}_2\text{O}/\text{Ni}^a$	number of functions	hydrogen shielding		oxygen shielding	
		orbital part	hyperfine part	orbital part	hyperfine part
aug-pcJ0/MK	200	30.4	-215.7	348.7	-17294.7
aug-pcJ1/MK	434	29.0	-243.4	357.0	-21139.9
aug-pcJ2/MK	848	28.8	-241.4	356.0	-20934.5
aug-pcJ0/MK+	248	30.4	-218.7	347.0	-17261.6
aug-pcJ1/MK+	482	29.0	-245.0	355.3	-21144.7
aug-pcJ2/MK+	896	28.8	-243.6	354.3	-20925.4

^a Notation: O and H basis/nickel basis. MK = Munzarová and Kaupp basis.²⁶ In MK+, partial decontraction is applied, and polarization functions are added (see text).

fleeing intermediate occurring in the dissociative exchange reaction in the FSS,²³ respectively. We were able to witness one such exchange process in our FPMD simulation,²³ which is longer than comparable first-principles studies hitherto performed.

We discuss the accuracy of the models based on a comparison among several available first-principles methods and with experimental data, as well as the general suitability of the FPMD trajectory for the static and dynamic magnetic properties in this system, together with the basis-set convergence. The spin density distribution and magnetic properties, HFC, \mathbf{g} , and ZFS tensors, as well as the pNMR shieldings, are reported.

RESULTS AND DISCUSSION

Detailed descriptions are included in the Supporting Information about the methods used for static models of the first solvation shell, FPMD simulations, as well as property calculations on simulation snapshots.

Basis Set Selection for the $\text{Ni}(\text{H}_2\text{O})_6^{2+}$ Complex. The results for the basis-set convergence for the calculation of NMR/ESR properties are listed in Table 1. The largest changes between results obtained with different hydrogen and oxygen basis sets are seen between aug-pcJ0 and aug-pcJ1, whereas aug-pcJ2 only brings a small further modification. Improving the MK basis for Ni by diffuse/polarization functions and partial decontraction (MK+) similarly produces only a slight change of results. Therefore, for production calculations, the aug-pcJ1/MK combination of basis sets was selected.

Methods for pNMR Shielding along the MD Trajectory. To assess the extent of the region that is significantly influenced by the paramagnetic center (Figure 1), we first tested the radial decay properties of the spin density at the PBE/SVP level. The spin density decays very rapidly around the metal ion, so that relatively little is left, on average, in SSS. On the basis of these data, we decided to include the complete FSS and SSS in our production calculations using the larger basis sets. Although SSS is relatively well-defined on average, in instantaneous configurations, its border to bulk water is not clear. We considered every water molecule that has at least one atom closer than 5 Å from the central ion, to belong to SSS. In this context, we noticed that the implicit solvation model (COSMO for the ESR tensors calculated with ORCA,²⁷ as well as PCM for the orbital shielding in G03²⁸) both improves the self-consistent field (SCF) convergence and prevents the occasional occurrence of spurious high spin density in the outer regions of the calculated snapshots

clusters. These implicit solvation models were therefore used in all further snapshot calculations. The explicit solvation of the selected SSS region was accomplished by including in the snapshot calculations all the water molecules that have at least one atom closer than 7 Å from the central atom. This layer corresponds roughly to the extent of the third solvation shell of the Ni^{2+} ion. This procedure resulted in clusters that contained on average 55.36 water molecules, of which 16.07 were in SSS, with FSS consisting most of the time of the six nearest neighbors of the ion. These numbers fluctuated in the ranges 48–63 and 10–22 for the total size of the cluster and SSS, respectively. To significantly decrease the computational cost, the small SVP basis was used beyond SSS. Our tests proved that a completely negligible effect resulted from using the SVP basis as compared to a calculation, in which the larger basis set was used throughout. A total of 2469 such clusters were sampled from the equilibrated part of the trajectory of the system containing 127 water molecules.

Structure of the Solvated Ni^{2+} Ion. The structural parameters for various combinations of methods are reported in a related work.²³ Briefly, in the six-fold structure, the oxygen atoms are octahedrally coordinated, whereas the symmetry of the whole complex is weakened by the increasing tilt angle of the water molecules, defined as the angle between the Ni–O vector and the HOH bisector. The 5-fold coordinated complex forms a tetragonal pyramid. The same is true for the coordination found in the corresponding parts of the FPMD simulation trajectory. As an example, frequently used in further calculations, the static calculations using PBE *in vacuo* and PBE together with the COSMO solvation model can be taken. The first results in a Ni–O distance of 2.086 Å and tilt angle of 23.0°, whereas the COSMO structure was 2.076 Å and 45.7°, the latter resembling the average FPMD structure.

Hyperfine Couplings in the $\text{Ni}(\text{H}_2\text{O})_6^{2+}$ and $\text{Ni}(\text{H}_2\text{O})_5^{2+}$ Complexes. To compare the impact of various exchange-correlation functionals (ECFs) on the isolated and hexa- and penta-coordinated dications, we selected a single structure obtained by geometry optimization at the PBE/def2-TZVP level. The Fermi contact term A_{con} , the largest (in absolute value) principal value of spin-dipole tensor A_{dip} (denoted as A_{33}^{dip}), as well as the rhombicity ($\text{Rh}(\mathbf{X}) = X_{22} - X_{11}$) $\text{Rh}(A_{\text{dip}})$ of the HFC tensor were calculated for both the ^1H and ^{17}O nuclei. One set of calculations was carried out *in vacuo*, i.e., no explicit or implicit solvation model was applied around $\text{Ni}(\text{H}_2\text{O})_6^{2+}$. The results are illustrated in Figures 2 and 3 for ^1H and ^{17}O , respectively. Both the A_{con} and A_{33}^{dip} parameters develop systematically with the exact exchange admixture in the

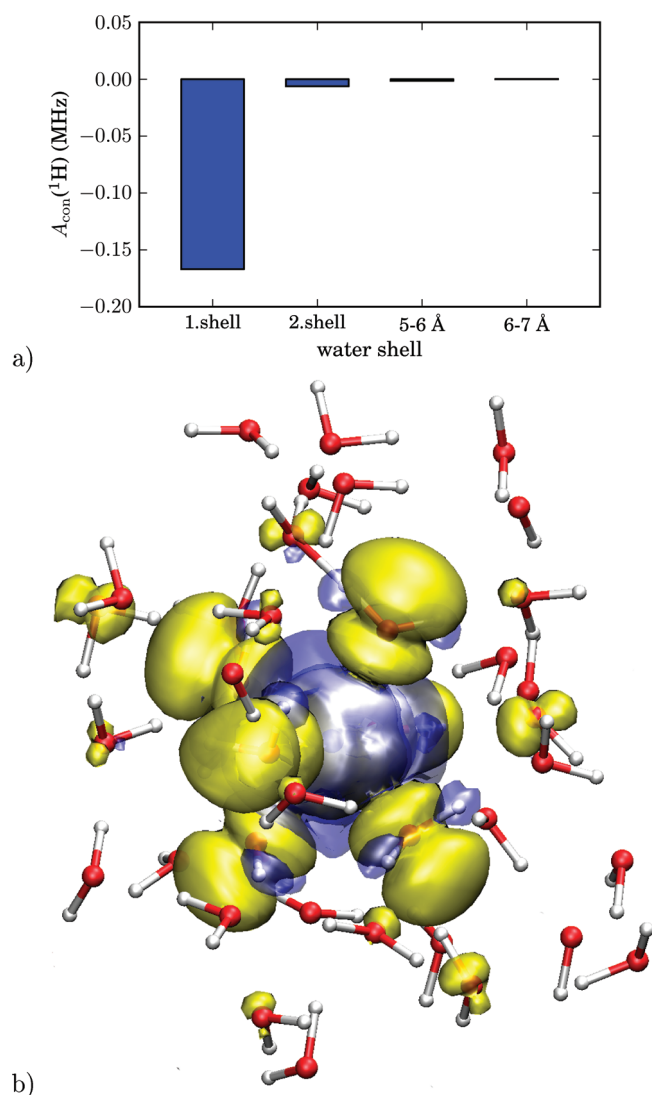


Figure 1. (a) Simulated Fermi contact hyperfine coupling constant at the ^1H nuclei in an aqueous solution of Ni^{2+} ion, as a function of increasing distance from the metal center. The presented values are averages over 10 randomly chosen snapshots calculated using the PBE/SVP level of theory, together with the COSMO solvation model. (b) Spin density in a random simulation snapshot. The yellow isosurface shows the positive spin density of 0.0001 au, whereas the blue isosurface indicates a negative spin density of the same absolute value. The spin density is taken from the periodic QUICKSTEP²⁹ calculation in which the PBE/DZVP level of theory was used.

ECF. Large relative differences are observed among the $A_{\text{con}}(^1\text{H})$. To produce these data, all structures were obtained with the def2-TZVP basis set, and for the property calculations, the aug-pcJ1/MK combination was used for O, H/Ni. Tabulated numerical data can be found in Tables S2 and S3 in the Supporting Information. The $A_{33}^{\text{dip}}(^1\text{H})$ exhibits a weaker, although equally systematic dependence on the choice of the ECF and is roughly one magnitude larger in absolute value. The situation is different for the ^{17}O nuclei, where A_{con} dominates over the spin-dipole part. The largest absolute value of A_{con} (−35.83 MHz for BLYP) is almost twice as large as the smallest among the tested set of methods (−20.40 MHz using SCS-MP2).

To test the direct impact of the solvation model on the calculated hyperfine properties, we used the same PBE/def2-TZVP/*in vacuo*

structure but employed the COSMO solvation model for the hyperfine calculation. In the scale of the differences among different methods (ECFs or SCS-MP2), the impact of the implicit solvation model in the property calculation step is entirely negligible. To test, on the other hand, the indirect impact of the COSMO solvation model on properties via its effect on the structure, we performed the property calculation also using PBE/def2-TZVP/COSMO structures. The latter differ significantly from the corresponding *in vacuo* structures, both in the Ni–O distance and tilt angle. The largest relative change is observed for $A_{\text{con}}(^1\text{H})$, proportional to the spin density on these nuclei, where not only the absolute value but even the sign changes. The influence of using the COSMO structure on the spin dipole contribution to ^1H and both $A_{\text{con}}(^{17}\text{O})$ and $A_{33}^{\text{dip}}(^{17}\text{O})$ is relatively less important but still numerically significant. The absolute values of the dipolar contributions for both ^1H and ^{17}O increase due to using the implicitly solvated structure, whereas $A_{\text{con}}(^{17}\text{O})$ decreases.

We have further consistently optimized all structures with the same functional (or SCS-MP2) as the one used for the property calculation and obtained data sets, both using COSMO and *in vacuo*. Subsequently, the properties were calculated only using, for simplicity, the COSMO model (the influence of COSMO in the property step is seen above to be minor). For the COSMO structures, $A_{\text{con}}(^1\text{H})$ and $A_{\text{con}}(^{17}\text{O})$ span a narrower range as compared to the two data sets obtained using the PBE structures. This implies that using structure optimization at the same level as used for the property calculation may be the preferred choice for hyperfine properties.³⁰

We have done all of the above calculations also for the 5-fold coordinated Ni^{2+} ion. The results are plotted to the same scale as for the 6-fold coordinated case, in Figures 2 and 3. Upon the change from 6- to 5-fold coordination, there is a clear increase of the magnitude of the dipolar contribution for the ^1H nuclei and a slight increase for the ^{17}O nuclei (in absolute value) in all of the calculated static structures, as well as in the average values over the FPMD trajectory. This observation is in accordance with the shortened distances between Ni^{2+} and the nuclei in question. For the contact term, only a very slight and nonuniform change takes place for the ^1H nuclei. For ^{17}O , we observe a slight decrease in the magnitude of A_{con} in 32 of the 36 tested static calculations, as well as in the FPMD average.

The overall conclusion is that the signs and orders of magnitude are well-defined across the different levels of theory used, for the presently relevant HFC parameters other than $A_{\text{con}}(^1\text{H})$. In light of these results, the agreement between the results from the experiment³¹ [$A_{\text{con}}(^1\text{H}) = -0.13$ MHz] and PBE (COSMO/COSMO for properties/structure; −0.15 MHz; Table S2) should be considered rather fortuitous. The results show that it is important to account for solvation effects on the underlying molecular structures.

g Tensor. The **g** tensor parametrizes the Zeeman interaction of the electronic spin with the magnetic field and enters the hyperfine terms of the pNMR shielding. In particular, the s.c. **g** shift tensor, $\Delta\mathbf{g} = \mathbf{g} - g_e\mathbf{1}$, i.e., the deviation from the isotropic free-electron **g** factor ($g_e \approx 2.0023$), is a nontrivial quantity to compute. In complete analogy to the test discussed above concerning HFC, we tested the performance of the available ECFs, the inclusion of a solvation model, as well as the dependence on the structure obtained by different methods. The results are graphically presented in Figure 4 and tabulated in Table S4, Supporting Information. We observe that also the **g** value is systematically dependent on the exact exchange admixture of the ECF, with the **g** value rising with an increasing fraction of the exact exchange. We lack presently efficient tools for

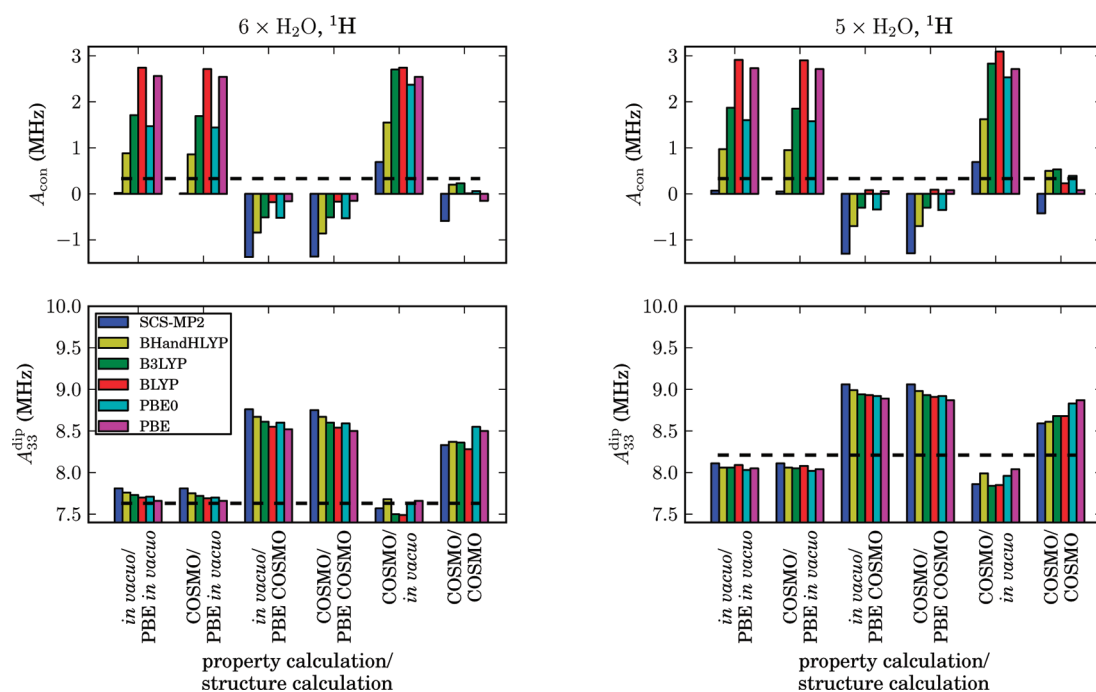


Figure 2. Impact of the choice of the DFT exchange-correlation functional on the average Fermi contact (con) and spin-dipole (dip) part of the ^1H hyperfine coupling (MHz) in $\text{Ni}(\text{H}_2\text{O})_6^{2+}$. The results of the SCS-MP2 calculations are presented for comparison. For the first four sets, the structures were obtained using the PBE functional, whereas for the last two sets, the structures were obtained with the same functional as the properties. All values present averages over all nuclei. Only the largest eigenvalue (in absolute value) A_{33} of the spin-dipole tensor is plotted. The dashed lines illustrate the average values obtained from the FPMD simulation (using PBE both for the trajectory and for the properties) of 127 water molecules and one Ni^{2+} ion. The corresponding (6- or 5-fold coordinated Ni^{2+} ion) parts of the trajectory were used for obtaining these values.

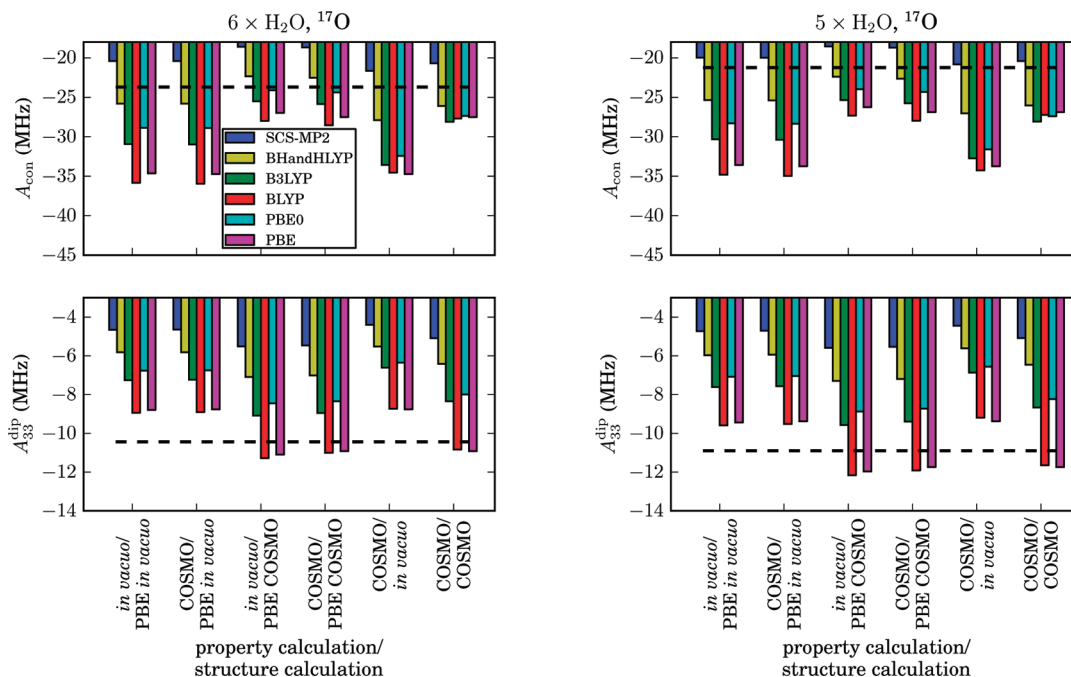


Figure 3. As for Figure 2, but for ^{17}O .

obtaining *ab initio* estimates for the g value in these complexes, although an implementation at the coupled-cluster singles and doubles (CCSD) level of theory has already been reported.³² Such tools would be highly useful for calibrating the DFT

performance, as no direct experimental results on the g tensor have been reported for the present system.

With any present ECF, the obtained g value is larger by 0.01–0.02 for the 5-fold than for the 6-fold coordinated ion.

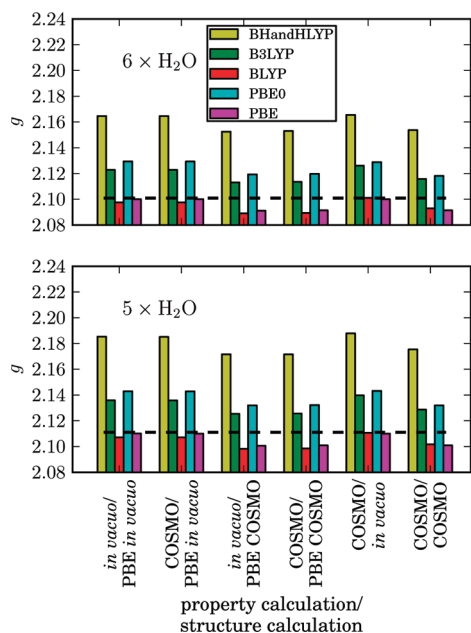


Figure 4. As for Figure 2, but for the isotropic g values.

Second, the g value is always smaller by roughly the same amount for the 5- and 6-fold coordinated structures obtained with the solvation model than for the corresponding *in vacuo* structures. The structures obtained with COSMO generally have a larger tilt angle and shorter Ni–O distance than structures obtained without the solvation model.²³ Similarly to the HFC tensor, the effect of using the implicit solvation model in the property calculation step is found to be negligible for the g tensor. In contrast to HFC, however, using the same ECF both for the structure and for the g tensor does not change the results appreciably, as compared to using PBE for structures.

Averages obtained from the MD snapshots using the PBE functional indicate that the combined dynamic and solvation influence on the g value is only 0.001 for both the six- and five-coordinated situations (compared to *in vacuo* static structures), indicating partial cancellation of the two effects. A notable feature apparent in the plot of the g value along the MD trajectory (Figure 5) is its increase in the last part of the trajectory, when the system enters the intermediate, 5-fold coordinated situation. In this part of the trajectory, the parallel component g_{\parallel} can be defined by the direction of the “missing” water molecule. This component remains close to the isotropic value of the 6-fold coordinated average, whereas the perpendicular component g_{\perp} is responsible for the increase of the isotropic value.

The simulated g values are higher by up to 0.01 than the results of all static FSS models, when the data obtained by the same (PBE) functional are compared. Furthermore, the simulation verifies the size of the increase of g from the 6-fold-coordinated to the 5-fold-coordinated model, obtained using the static $\text{Ni}(\text{H}_2\text{O})_n^{2+}$ ($n = 5, 6$) structures.

Experimental data are available for the crystals of nickel Tutton salts, which contain Ni^{2+} coordinated by six water molecules.³³ The average g value for these compound is 2.25. The same value was later adopted also in an EPR relaxation study of the aqueous solution of Ni^{2+} .³⁴ The values of 2.25 and 2.3 have been also used in NMR studies of aqueous Ni^{2+} solutions (refs 35 and 31, respectively). Our static DFT calculations using different ECFs

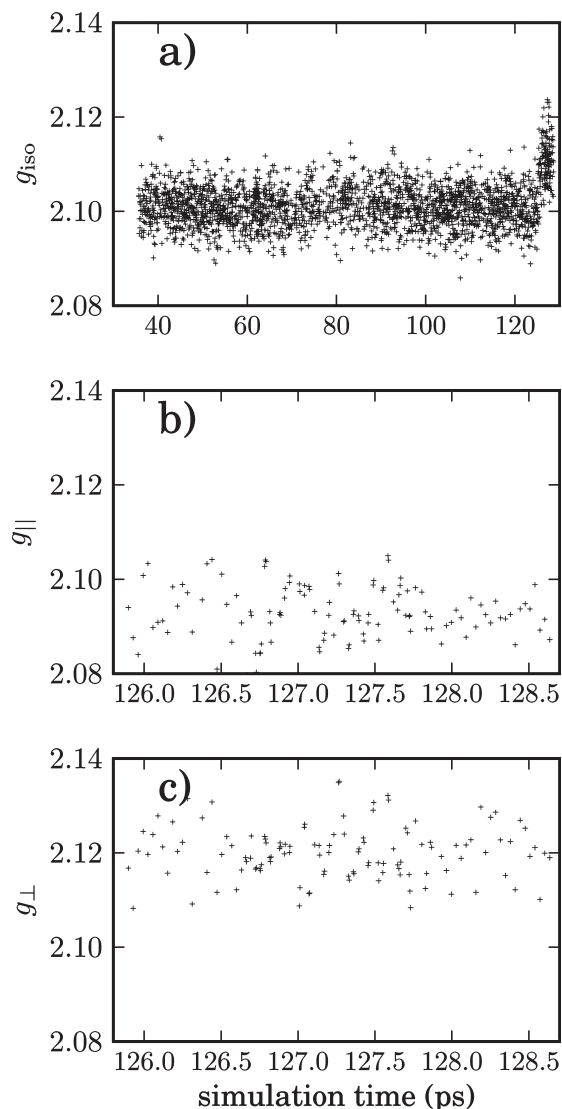


Figure 5. Calculated g tensor plotted along a first-principles molecular dynamics trajectory of the Ni^{2+} ion in the aqueous solution. The increase at the end of the trajectory (a) corresponds to a situation with a 5-fold coordinated Ni^{2+} ion. In panels b and c, parallel and average perpendicular eigenvalues are plotted for the 5-fold coordinated part of the trajectory. Here, g_{\parallel} and g_{\perp} are defined as the components of the g tensor along with and perpendicular to, respectively, the direction from the Ni^{2+} ion to the “missing” water molecule in the originally octahedral complex. The PBE functional was used for the system with 127 water molecules.

as well as the dynamic simulation point to significantly smaller g values. This implies that the assumption that the g value is the same in both salt crystal and liquid solution is questionable.

Zero-Field Splitting. The effect of the ZFS Hamiltonian

$$H_{\text{ZFS}} = \mathbf{S} \cdot \mathbf{D} \cdot \mathbf{S} \quad (1)$$

on the energy levels within the spin manifold of an $S \geq 1$ species is best characterized with parameters

$$D = D_{33} - \frac{1}{2}(D_{11} + D_{22}) \quad (2)$$

$$E = \frac{1}{2}(D_{22} - D_{11}) \quad (3)$$

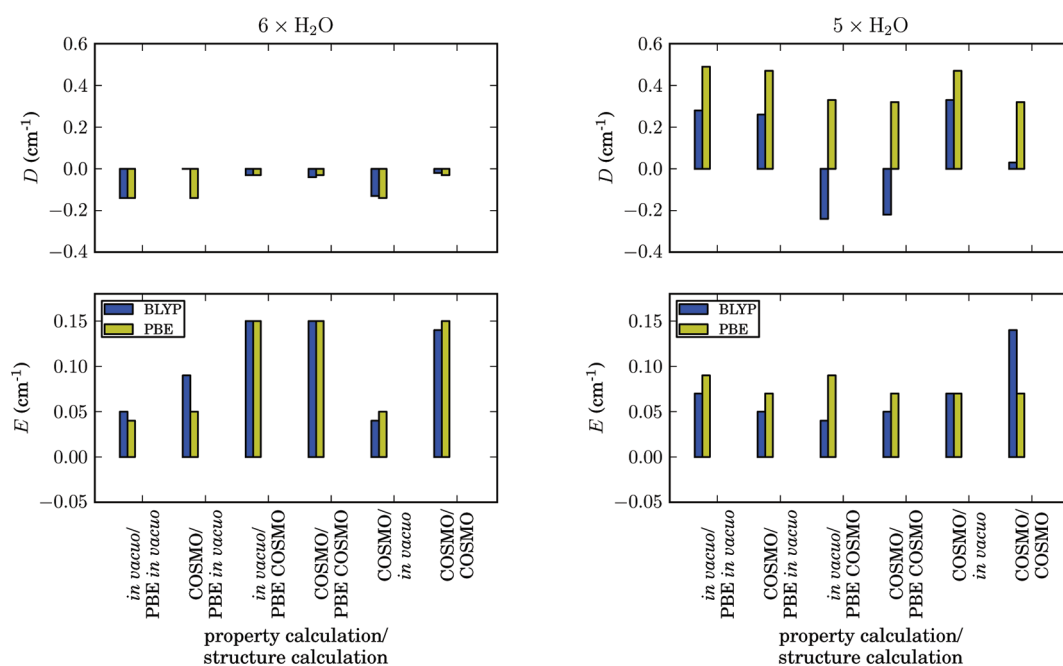


Figure 6. As for Figure 2, but the D and E parameters of the zero-field splitting tensors are presented. In all calculations, the same selection of the principal axis system defining the D and E parameters [eqs 2–3] was used so that similar geometries have similar orientations of the ZFS axis system.

expressed in terms of the eigenvalues D_{ii} of the \mathbf{D} tensor. The ordering of the principal values is selected so that $0 \leq E/D \leq 1/3$.^{36,37}

In the calculation of the ZFS tensor, we tested two GGA functionals, PBE and BLYP, the inclusion/omission of the solvation model as well as the dependence on the structure obtained by different methods. The same calculations were carried out for both $\text{Ni}(\text{H}_2\text{O})_n^{2+}$ ($n = 5, 6$) complexes. The results are plotted in Figure 6, and the numerical data are contained in Table S5, Supporting Information. In the upper right panel of Figure 6, the discrepancy between BLYP and PBE functionals when COSMO structures are used is caused by the poor alignment between the principal axis of the ZFS tensor and the molecular axis system. The negative sign in the two cases of the D parameter determined by BLYP is therefore not entirely relevant. Since the structure of the 6-fold coordinated Ni^{2+} ion is highly symmetric, the resulting D and E values are expectedly close to zero. The symmetry is broken in the 5-fold coordinated ion.

For the calculations of snapshots extracted from the FPMD dynamics, it is important to consider the choice of the axis system. For any instantaneous configuration, a proper axis system can be found such that the D and E parameters can be calculated in the same way as described above. The time evolution of these parameters at the PBE level is plotted in Figure S1, Supporting Information. Clearly, both D and E vanish on average for an isotropic solution of the symmetrically coordinated ion. In this case, there is no choice of the axis system that would be unique and constant over time. For the 5-fold coordinated part of the trajectory, similarly as for the \mathbf{g} tensor, the parallel component D_{\parallel} can be defined by the direction from the nickel ion to the “missing” water molecule. The D and E parameters of the ZFS tensor are then given by eqs 2 and 3 with $D_{33} = D_{\parallel}$, D_{11} and D_{22} being the perpendicular components. The RMS values of D , E , and E/D can be found for

the 5-fold-coordinated part of the trajectory in Table 2. For the 6-fold-coordinated part, which is on average octahedrally symmetric, the RMS value of the scalar tensor product (Δ)

$$\Delta = \sqrt{\langle \mathbf{D}(0) : \mathbf{D}(0) \rangle} = \sqrt{\left\langle \sum_{\alpha\beta} D_{\alpha\beta}(0) D_{\alpha\beta}(0) \right\rangle} \quad (4)$$

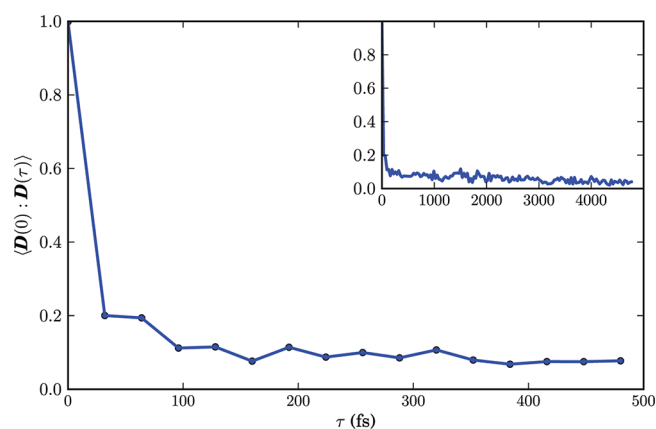
serves as a parameter characterizing the fluctuations of the ZFS tensor. For comparison, we also performed an analysis equivalent to that of Odelius et al.¹¹ and calculated the time-correlation function (TCF) of the scalar tensor product $\langle \mathbf{D}(0) : \mathbf{D}(\tau) \rangle$ (TCF_{ZFS}). The normalized autocorrelation function is plotted in Figure 7. The value of the normalized TCF decays to less than 0.1 in a little over 100 fs and slowly decreases to the noise level during the following ca. 4 ps.

In our simulation, the decay of the TCF does not exhibit the well-defined damped oscillatory character seen by Odelius et al. with a period of roughly 90 fs. In their work, they obtain the instantaneous ZFS tensor from a preparameterized function involving the geometrical distortions of the octahedral complex along two classes of modes corresponding to Ni–O bond vibrations (E_g modes) and oscillation of the O–Ni–O angle. The dependence on other modes, namely, the rotational mode of the water molecules, was omitted. The oscillatory TCF of the E_g modes is almost identical to TCF_{ZFS} . It may be speculated that the remaining geometrical dependences cause the TCF_{ZFS} to be more damped. According to our calculation, the ZFS tensor is more sensitive to molecular geometry than what the model of Odelius et al. assumed. For example, the $\langle \mathbf{D}(0) : \mathbf{D}(\tau) \rangle$ curve from our work is damped to an extent that no anticorrelation can be observed. Also, the calculated RMS value Δ yields 100 cm^{-1} in our study for the 6-fold coordinated part of the FPMD trajectory, which is an order of magnitude higher than that of Odelius et al. ($\sim 5.2 \text{ cm}^{-1}$).¹¹ The reason for such a large difference is presently unclear to us.

Table 2. Calculated EPR and NMR Parameters of the Aqueous Solution of Ni²⁺^a

	6 × H ₂ O ^b	5 × H ₂ O ^c
g_{iso}^d	2.1009 ± 0.0002	2.1110 ± 0.0004
$g_{\text{iso}}(\text{exptl})^e$	2.25	
RMS(D) (cm ⁻¹) ^f		14.6
RMS(E) (cm ⁻¹) ^g		5.1
RMS(E/D)		9.1
Δ (cm ⁻¹) ^h	100	13.9
Δ_5 (cm ⁻¹) ⁱ	45	6.2
Δ_5 (cm ⁻¹ ; exptl) ^j	2.6	
$A_{\text{con}}(^1\text{H})$	0.33 ± 0.07	0.33 ± 0.07
$A_{\text{con}}(^{17}\text{O})$	-23.7 ± 0.3	-21.2 ± 0.4
$A_{\text{dip},33}(^1\text{H})^k$	7.63 ± 0.02	8.21 ± 0.03
$A_{\text{dip},33}(^{17}\text{O})^k$	-10.44 ± 0.05	-10.9 ± 0.2
Rh[$A_{\text{dip}}(^1\text{H})$] ^l	2.73 ± 0.02	3.06 ± 0.02
Rh[$A_{\text{dip}}(^{17}\text{O})$] ^l	0.090 ± 0.002	0.38 ± 0.09
$A_{\text{iso}}(^1\text{H})^m$	0.30 ± 0.07	0.29 ± 0.07
$A_{\text{iso}}(^{17}\text{O})^m$	-23.5 ± 0.3	-20.9 ± 0.4
$A_{\text{iso}}(^1\text{H})(\text{exptl})$	0.13 ± 0.01 ⁿ	
$A_{\text{iso}}(^{17}\text{O})(\text{exptl})$	24, ^o -28.2 ^p	
$\sigma_{350\text{K}}^{\text{iso}}(^1\text{H})$	6 ± 4	
$\sigma_{350\text{K}}^{\text{iso}}(^{17}\text{O})$	-10960 ± 90	
$\sigma_{350\text{K}}^{\text{iso}}(^1\text{H})(\text{exptl})$	34 ^q	
$\sigma_{350\text{K}}^{\text{iso}}(^{17}\text{O})(\text{exptl})$	-10988, ^r -9938 ^s	

^a First principles molecular dynamics using the PBE functional and 127 water molecules. Experimental results are listed where available. ^b The six-coordinated part of the simulation trajectory, statistics over 2000 snapshots. ^c The five-coordinated part of the simulation trajectory, statistics over 120 snapshots. ^d The corresponding g shift value would be obtained by subtracting the free-electron value $g_e = 2.0023$. ^e Ref 33. Measurement on hydrated Ni²⁺ salt crystals. This value has, however, been used also for Ni²⁺ solutions in the literature. ^f For the five-coordinated complex, the parallel axis is in the direction of the “missing” water molecule in the originally octahedral complex; then, $D = D_{\parallel} - 1/2(D_{\perp,1} + D_{\perp,2})$. For the 6-fold coordinated complex, these quantities are not clearly defined (see text). ^g As footnote f , but $E = 1/2(D_{\perp,1} - D_{\perp,2})$. ^h RMS obtained from the average scalar tensor product eq 4. ⁱ $\Delta_5 = (\langle \mathbf{D}(0) : \mathbf{D}(0) \rangle / 5)^{1/2}$ (here, the factor 5 is the number of independent components needed to specify the ZFS tensor). The numbers are added for consistency with the literature experimental value (footnote j). ^j Ref 34. Electron spin relaxation study based on proton NMR relaxation measurement in aqueous Ni²⁺ solution. ^k The largest (in absolute value) value of the eigenvalues of the traceless tensor. ^l Rhombicity of the dipolar tensor, $\text{Rh}(A_{\text{dip}}) = A_{22}^{\text{dip}} - A_{11}^{\text{dip}}$. The values are very small for the oxygen nuclei in the 6-fold-coordinated complex, in agreement with the nearly octahedral symmetry. ^m The isotropic value contains both the Fermi contact and spin-orbit contributions, the latter being much smaller. ⁿ Ref 31. The hyperfine coupling constant was reported without sign, an inspection of the data points to the negative sign. ^o Ref 35. The hyperfine coupling constant was reported without sign. ^p Ref 38. ^q This value of the shielding constant was obtained by extrapolation of the experimental data measured in the range of 243.15–263.15 K³¹ assuming a $1/T$ dependence of the paramagnetic chemical shifts. Furthermore, pure water in gas phase (PBE) was taken as the reference, $\sigma_{\text{ref,gas}}(^1\text{H}) = 31.59$ ppm with a gas-to-liquid shift correction (B3LYP) $\delta_{\text{gas} \rightarrow \text{liquid}}(^1\text{H}) = -5.27$ ppm,⁶ see the main text. ^r The value of the shielding constant was obtained by extrapolation of the experimental data measured in the range of 274.65–306.15 K³⁵ assuming a $1/T$ dependence of the paramagnetic chemical shifts. Referenced and corrected for gas-to-liquid shift as in footnote q using the values for ¹⁷O: $\sigma_{\text{ref,gas}}(^{17}\text{O}) = 324.8$ ppm, $\delta_{\text{gas} \rightarrow \text{liquid}}(^{17}\text{O}) = -41.2$ ppm.⁶ From the experimental data measured in the range of 243.65–308.15 K.³⁸ Referenced and corrected for the gas-to-liquid shift as in footnote r .

**Figure 7.** Simulated time correlation function $\langle \mathbf{D}(0) : \mathbf{D}(\tau) \rangle$ of the scalar tensor product of the zero-field splitting tensor of the aqueous solution of Ni²⁺. First-principles molecular dynamics using the PBE functional and 127 water molecules.

Nuclear Shielding in the Static Structures. We present calculated shielding contributions using the static 5- and 6-fold-coordinated Ni(H₂O)_{*n*}²⁺ optimized both *in vacuo* and with the COSMO implicit solvation model. We use the classification and nomenclature of the terms set forth in the general shielding theory for species of arbitrary spin multiplicity, in ref 25. In particular, and as described in the Supporting Information, the theory predicts for a higher-than-doublet system several new isotropic shielding contributions as compared to the case of the doublet.³⁹ In the present triplet system, however, the influence of these new terms turns out to be small.

In the calculation of the nuclear shielding contributions, in analogy to the tests of the HFC, \mathbf{g} , and ZFS tensors, we have tested the performance of the present set of ECFs, inclusion/omission of the solvation model, and the dependence on the structure obtained by different methods. The results are summarized in Figures 8 and 9, and the numerical data are given in Tables S6–S9, Supporting Information.

The observations at different levels of theory are parallel to the corresponding experience with HFC. The total isotropic shielding is dominated by the contact term for both the ¹H and ¹⁷O nuclei. For ¹H, there is a major influence of the structure at which the shielding tensor is calculated. For the structures obtained *in vacuo*, the ¹H contact shieldings have a sign opposite that for the structures obtained with the implicit COSMO solvent model. In the case of ¹⁷O, the use of COSMO structures leads to a reduction of the contact term by 1000–3000 ppm.

For all of the terms, there is a clear dependence of the observed value on the amount of the exact exchange admixture. This is particularly important for the dominating contact term. In contrast, all of the hyperfine shielding terms exhibit only a minor dependence on the use of the COSMO model at the property calculation step. The solvent effect arises mainly indirectly, via the structure. The orbital shielding σ_{orb} changes, however, also upon the introduction of the PCM solvent model in the property calculation step.

Apart from the nonrelativistic σ_{con} and σ_{orb} contributions, the isotropic shielding obtains contributions from $\sigma_{\text{con},2}$ and $\sigma_{\text{con},3}$ terms arising from the relativistic spin-orbit contributions to the HFC and \mathbf{g} tensors, respectively. None of the other terms contribute numerically significantly. In particular, both the pseudocontact and the novel isotropic terms introduced in

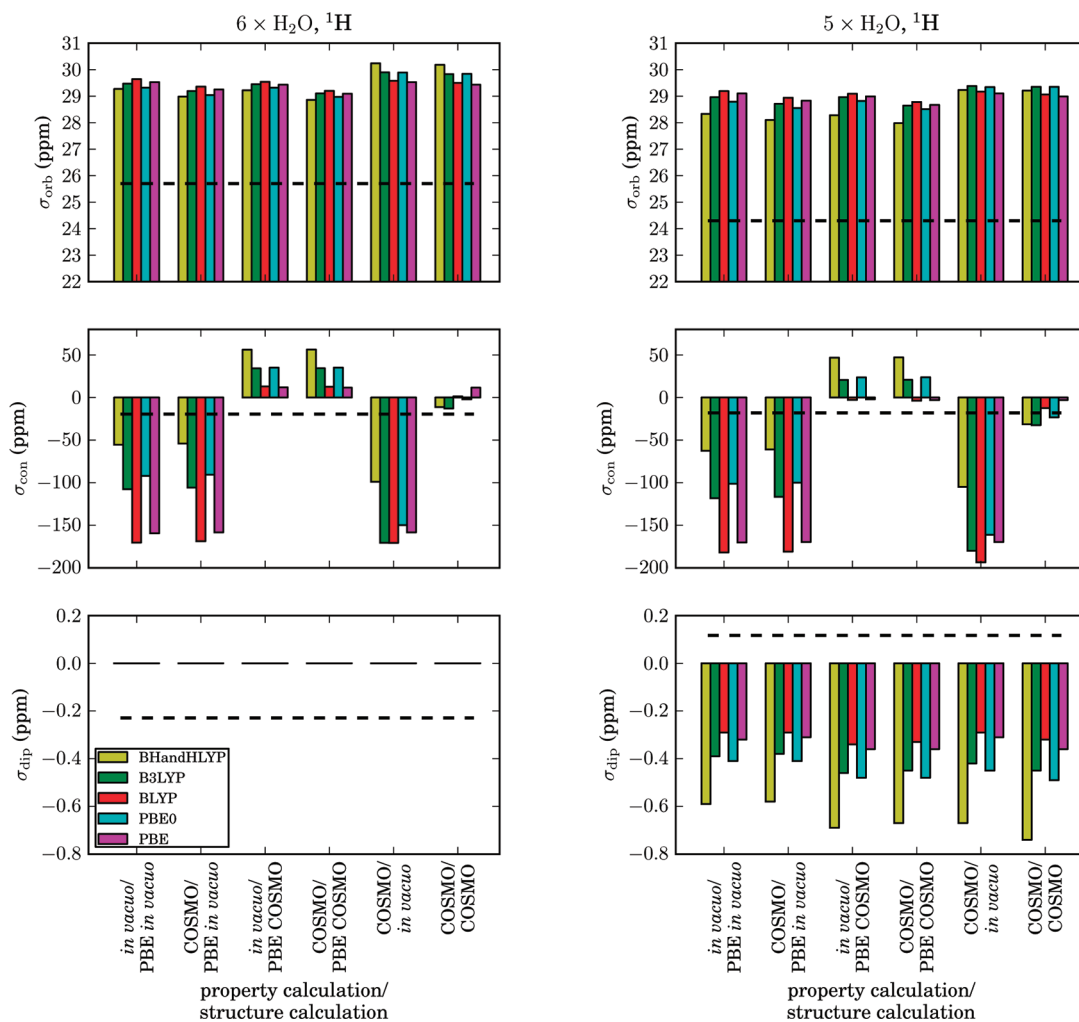


Figure 8. Impact of the choice of the DFT exchange-correlation functional on the average orbital (orb), contact (con), and dipole (dip) parts of the ^1H isotropic shielding constant (ppm) in $\text{Ni}(\text{H}_2\text{O})_n^{2+}$ ($n = 5, 6$). Here, σ_{con} consists of the sum of isotropic values of $\sigma_{\text{con},1}$, $\sigma_{\text{con},2}$, and $\sigma_{\text{con},3}$ terms; σ_{dip} is calculated in an analogous way by summing isotropic values of $\sigma_{\text{dip},1}$, $\sigma_{\text{dip},2}$, $\sigma_{\text{dip},3}$, and σ_{pc} (Table S1, Supporting Information).

ref 25 for higher than doublet spin multiplicities play an unimportant role in the present triplet system.

The smallest range of the total σ is again observed for the structures obtained with the COSMO model and using the same functional by which the property calculations were performed. Due to the ambiguity of the $A_{\text{con}}(^1\text{H})$, the final $\sigma(^1\text{H})$ in FSS remains somewhat vaguely determined, in the range of 15–39 ppm. The relative range pertaining to ^{17}O is smaller, and the final results are between -11200 and -10500 ppm.

Reducing the FSS of the Ni^{2+} ion from six to five water molecules, a decrease is observed for the orbital shielding of the ^1H nuclei (Figure 8, up-most panels), whereas an increase is seen for ^{17}O for all of the tested combinations of static structures and functionals, as well as for the averages from the FPMD trajectory (Figure 9, up-most panels). The sum of the isotropic contact contributions σ_{con} follows the behavior of A_{con} . Hence, there is only very little change at ^1H nuclei and decrease in magnitude for ^{17}O .

The sum of the isotropic dipolar contributions ($\sigma_{\text{dip},1}$, $\sigma_{\text{dip},2}$, $\sigma_{\text{dip},3}$, and σ_{pc}) gives a negligible average for the highly symmetric 6-fold-coordinated structures because of the negligible differences in the ZFS energy levels. Nonzero values are obtained for

the 5-fold coordinated structures, which reflects the behavior of the ZFS tensor. Among these terms, the σ_{pc} term strongly dominates for ^1H in 5-fold situations. For the ^{17}O nucleus, the magnitudes of σ_{dip} and σ_{pc} are similar for the 6-fold coordinated structure, whereas the σ_{pc} term dominates in the 5-fold structure. Also, from the FPMD trajectory, small but nonvanishing average values of the isotropic dipolar shielding are obtained for both nuclei.

Nuclear Shielding in MD Snapshot Calculations. An analysis of the nuclear shielding terms obtained from the FPMD simulation is presented in Tables 3 and 4. The same orbital and contact terms are seen to give significant contributions in the dynamic solvation data as for the static structures. We again observe a dominant contact contribution to the isotropic pNMR shielding for both ^1H and ^{17}O nuclei, in both FSS and SSS. For both ^1H and ^{17}O in FSS, the second largest contribution is due to the con,3 term originating from the SO correction to \mathbf{g} . The con,2 term, which is related to the SO correction of \mathbf{A} , also contributes significantly. It is noteworthy that the dynamic FSS contact shielding of -20 ppm differs clearly from the corresponding static models of FSS (Table S6, Supporting Information), where the use of the COSMO model consistently

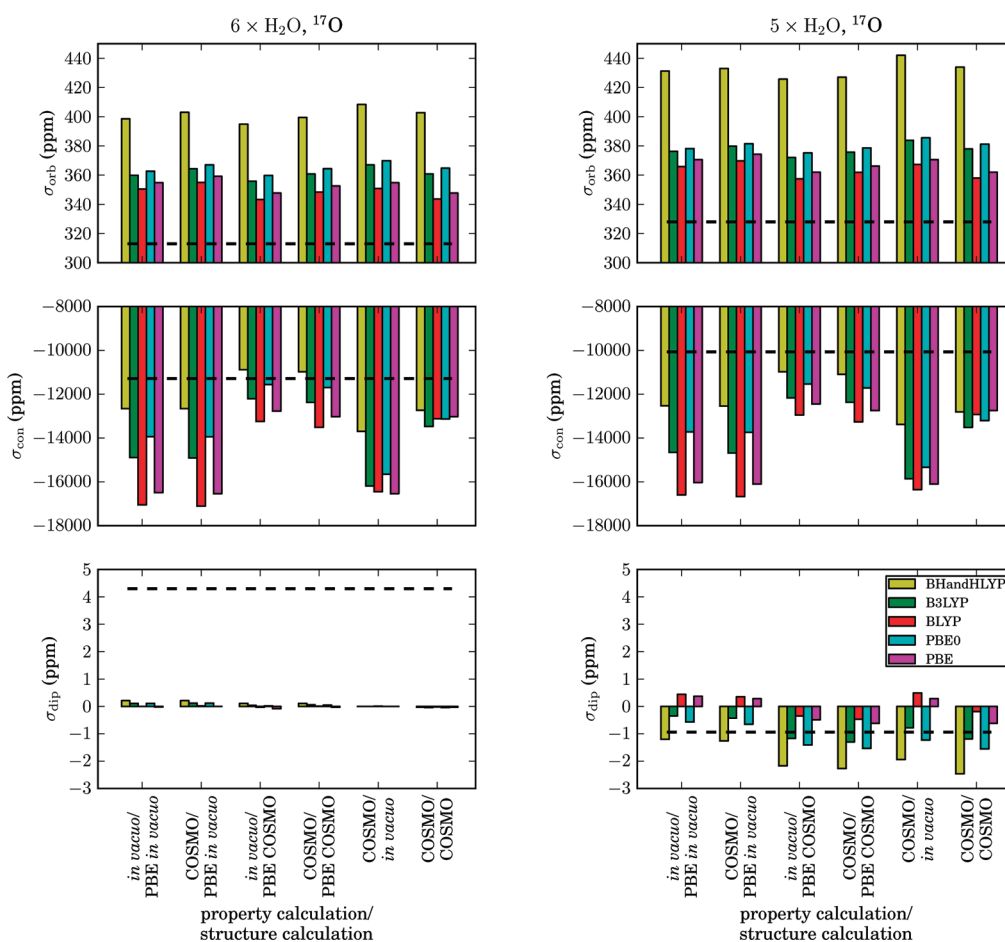


Figure 9. As for Figure 8, but for ^{17}O .

Table 3. Total Isotropic ^1H and ^{17}O Nuclear Magnetic Resonance Shielding Constants, Error Margins, and Contributions from Different Physical Mechanisms for Atoms in the First and Second Solvation Shells of the Aqueous Solution of the Ni^{2+} Ion^a

symbol	term in $\sigma_{\epsilon\tau}^b$	atom (shell)			
		^1H (FSS)	^1H (SSS)	^{17}O (FSS)	^{17}O (SSS)
σ_{orb}	$\sigma_{\epsilon\tau}^{\text{orb}}$	25.7 ± 0.1	25.64 ± 0.07	313 ± 1.0	269.2 ± 0.7
σ_{con}	$g_e A_{\text{con}} \langle S_{\epsilon} S_{\tau} \rangle_0$	-20 ± 4	0.4 ± 0.1	-10870 ± 90	-86 ± 8
σ_{dip}	$g_e \sum_b A_{b\tau}^{\text{dip}} \langle S_{\epsilon} S_b \rangle_0$	-0.18 ± 0.09	-0.04 ± 0.03	4 ± 1	-0.08 ± 0.02
$\sigma_{\text{con},2}$	$g_e A_{\text{PC}} \langle S_{\epsilon} S_{\tau} \rangle_0$	1.57 ± 0.01	0.0145 ± 0.0005	121.3 ± 0.2	-0.68 ± 0.02
$\sigma_{\text{dip},2}$	$g_e \sum_b A_{b\tau}^{\text{dip},2} \langle S_{\epsilon} S_b \rangle_0$	0.002 ± 0.004	-0.002 ± 0.001	0.11 ± 0.04	-0.004 ± 0.001
$\sigma_{\text{con},3}$	$\Delta g_{\text{iso}} A_{\text{con}} \langle S_{\epsilon} S_{\tau} \rangle_0$	-1.0 ± 0.1	0.022 ± 0.003	-534 ± 4	-4.28 ± 0.3
$\sigma_{\text{dip},3}$	$\Delta g_{\text{iso}} \sum_b A_{b\tau}^{\text{dip}} \langle S_{\epsilon} S_b \rangle_0$	-0.008 ± 0.003	-0.002 ± 0.001	0.19 ± 0.04	-0.004 ± 0.001
$\sigma_{\text{c,aniso}}$	$A_{\text{con}} \sum_a \Delta \tilde{g}_{\epsilon a} \langle S_a S_{\tau} \rangle_0$	-0.0002 ± 0.0003	0.0 ± 0.0	-0.06 ± 0.06	0.0 ± 0.0
σ_{pc}	$\sum_{ab} \Delta \tilde{g}_{\epsilon a} A_{b\tau}^{\text{dip}} \langle S_a S_b \rangle_0$	-0.043 ± 0.005	0.0 ± 0.0	0.00 ± 0.03	-0.001 ± 0.002
σ_{tot}		6 ± 4	26.1 ± 0.1	-10960 ± 90	177 ± 8

^aThe numbers are calculated from the 2000 snapshots (sampled over ca. 80 ps) of the first-principles molecular dynamics trajectory using the PBE functional with 127 water molecules and one Ni^{2+} ion. In this part of the trajectory, the Ni^{2+} ion is six-fold-coordinated. Paramagnetic shielding was obtained using a formal temperature of 350 K. ^bReferring to the tensor component combinations of \mathbf{g} and \mathbf{A} that give rise to the listed terms. See ref 25 and the brief account of theory in the Supporting Information. See Table 2 for comparison with available experimental results.

for both structures and properties leads to ca. +9 ppm at the PBE level. Similarly as for A_{con} , $\sigma_{\text{con}} = +9$ ppm coincides well with the experimental value (7.68 ppm). [This number was obtained by assuming that the measured value is contact shift referenced with respect to the orbital shielding of pure water (*vide infra*).]

Taking into account the pronounced dependence of $A_{\text{con}}(^1\text{H})$ on the used ECF, this is probably fortuitous. A similar observation applies to the orbital shielding, where the static model produces a value some 4 ppm larger than the present FPMD result, 25.7 ppm. These differences indicate that dynamics need to be

Table 4. As for Table 3 but for the Five-Fold-Coordinated Part of the Trajectory^a

symbol	term in $\sigma_{\varepsilon\tau}$	atom (shell)			
		¹ H (FSS)	¹ H (SSS)	¹⁷ O (FSS)	¹⁷ O (SSS)
σ_{orb}	$\sigma_{\varepsilon\tau}^{\text{orb}}$	24.3 ± 0.9	25.0 ± 0.3	328 ± 6	265 ± 3
σ_{con}	$g_{\varepsilon}A_{\text{con}}\langle S_{\varepsilon}S_{\tau} \rangle_0$	-19 ± 3	0.66 ± 0.05	-9709 ± 130	-56 ± 4
σ_{dip}	$g_{\varepsilon}\Sigma_b A_{b\tau}^{\text{dip}}\langle S_{\varepsilon}S_b \rangle_0$	0.08 ± 0.05	-0.025 ± 0.007	-0.3 ± 0.9	-0.008 ± 0.008
$\sigma_{\text{con},2}$	$g_{\varepsilon}A_{\text{PC}}\langle S_{\varepsilon}S_{\tau} \rangle_0$	2.01 ± 0.04	0.025 ± 0.004	165 ± 1	-0.72 ± 0.08
$\sigma_{\text{dip},2}$	$g_{\varepsilon}\Sigma_b A_{b\tau}^{\text{dip},2}\langle S_{\varepsilon}S_b \rangle_0$	0.013 ± 0.006	0.0 ± 0.0	0.17 ± 0.06	0.0 ± 0.0
$\sigma_{\text{con},3}$	$\Delta g_{\text{iso}}A_{\text{con}}\langle S_{\varepsilon}S_{\tau} \rangle_0$	-1.1 ± 0.15	0.036 ± 0.003	-525 ± 7	-3.1 ± 0.2
$\sigma_{\text{dip},3}$	$\Delta g_{\text{iso}}\Sigma_b A_{b\tau}^{\text{dip}}\langle S_{\varepsilon}S_b \rangle_0$	0.004 ± 0.003	-0.0014 ± 0.0004	-0.01 ± 0.05	-0.0005 ± 0.0009
$\sigma_{\text{c,aniso}}$	$A_{\text{con}}\Sigma_a \Delta g_{\varepsilon a}\langle S_a S_{\tau} \rangle_0$	-0.0006 ± 0.0002	0.0 ± 0.0	-0.19 ± 0.09	0.0 ± 0.0
σ_{pc}	$\Sigma_{ab} \Delta g_{\varepsilon a} A_{b\tau}^{\text{dip}}\langle S_a S_b \rangle_0$	0.02 ± 0.02	0.009 ± 0.004	-0.8 ± 0.5	-0.003 ± 0.006
σ_{tot}		6 ± 3	25.7 ± 0.3	-9733 ± 130	204 ± 4

^a 120 snapshots over ca. 3 ps were used.Table 5. Principal Values of the Total ¹H and ¹⁷O Shielding Tensors and the Contributions of the Various Physical Mechanisms for Nuclei in the First (FSS) and Second (SSS) Solvation Shells in the Aqueous Solution of the Ni²⁺ Ion^a

	¹ H (FSS)			¹⁷ O (FSS)		
σ_{orb}	21.9 ± 0.7	32.1 ± 0.9	20.6 ± 0.6	230 ± 7	317 ± 10	365 ± 12
σ_{con}	-29 ± 5	-29 ± 5	-28 ± 5	-11200 ± 200	-11300 ± 200	-10200 ± 200
σ_{dip}	-462 ± 3	149.4 ± 1.4	312 ± 2	-4590 ± 50	2390 ± 40	2200 ± 50
$\sigma_{\text{con},2}$	1.63 ± 0.03	1.59 ± 0.03	1.65 ± 0.03	128 ± 2	129 ± 2	116 ± 3
$\sigma_{\text{dip},2}$	-24.0 ± 0.2	13.7 ± 0.2	10.29 ± 0.14	-171 ± 3	59 ± 6	112 ± 7
σ_{ac}	0.006 ± 0.008	-0.007 ± 0.007	0.001 ± 0.005	-0.07 ± 0.09	0.01 ± 0.08	0.06 ± 0.04
$\sigma_{\text{con},3}$	-1.4 ± 0.3	-1.4 ± 0.2	-1.4 ± 0.3	-557 ± 8	-563 ± 8	-506 ± 10
$\sigma_{\text{dip},3}$	-23.0 ± 0.2	7.45 ± 0.09	15.57 ± 0.14	-229 ± 3	119 ± 2	110 ± 3
$\sigma_{\text{c,aniso}}$	0.011 ± 0.013	-0.038 ± 0.014	0.03 ± 0.02	-1.1 ± 1.3	1 ± 2	-1 ± 2
σ_{pc}	-0.07 ± 0.02	0.024 ± 0.009	-0.09 ± 0.03	-0.4 ± 0.5	-0.1 ± 0.5	0.4 ± 0.4
σ_{tot}	-515 ± 6	174 ± 5	331 ± 6	-16400 ± 200	-8900 ± 200	-7800 ± 200
	¹ H (SSS)			¹⁷ O (SSS)		
σ_{orb}	22.5 ± 0.8	18.8 ± 0.8	35.2 ± 1.3	254 ± 7	251 ± 7	276 ± 8
σ_{con}	0.62 ± 0.10	0.62 ± 0.10	0.68 ± 0.10	-82 ± 5	-81 ± 5	-78 ± 5
σ_{dip}	-99.3 ± 0.9	48.6 ± 0.5	50.7 ± 0.7	-146 ± 2	58.9 ± 1.3	87 ± 1
$\sigma_{\text{con},2}$	0.016 ± 0.002	0.016 ± 0.002	0.017 ± 0.002	-0.70 ± 0.03	-0.69 ± 0.03	-0.69 ± 0.03
$\sigma_{\text{dip},2}$	-5.16 ± 0.07	2.55 ± 0.04	2.61 ± 0.04	-4.06 ± 0.07	2.32 ± 0.05	1.73 ± 0.06
σ_{ac}	0.003 ± 0.002	-0.002 ± 0.002	-0.0002 ± 0.0007	0.000 ± 0.003	-0.004 ± 0.002	0.004 ± 0.002
$\sigma_{\text{con},3}$	0.031 ± 0.005	0.031 ± 0.005	0.034 ± 0.005	-4.1 ± 0.3	-4.0 ± 0.2	-3.9 ± 0.2
$\sigma_{\text{dip},3}$	-4.96 ± 0.06	2.43 ± 0.03	2.53 ± 0.04	-7.31 ± 0.14	2.94 ± 0.07	4.37 ± 0.09
$\sigma_{\text{c,aniso}}$	-0.0004 ± 0.0002	0.0000 ± 0.0002	0.0006 ± 0.0003	-0.024 ± 0.011	0.006 ± 0.011	0.011 ± 0.011
σ_{pc}	-0.003 ± 0.005	-0.011 ± 0.005	0.018 ± 0.008	-0.022 ± 0.008	-0.004 ± 0.007	-0.001 ± 0.007
σ_{tot}	-86.2 ± 1.4	73.0 ± 0.9	91.8 ± 1.1	10 ± 10	230 ± 9	287 ± 9

^a See also Tables 3 and 4. Since there is no unique principal axis frame of the shielding tensors, which would be preserved through the simulation time for a particular molecule, the averages are calculated from principal values sorted according to the magnitude of the instantaneous principal values of the total shielding in the order $\sigma_{11}^{\text{tot}} \leq \sigma_{22}^{\text{tot}} \leq \sigma_{33}^{\text{tot}}$.

accounted for in proton shielding in FSS; i.e., the static models are not capable of yielding quantitative results. In contrast, the ¹⁷O shielding shows a rather modest, systematic shift of ca. 1700 ppm (13%) for contact and 34 ppm (10%) for orbital shielding, for the 6-fold-coordinated Ni²⁺ ion. The ¹H nuclei are expectedly more sensitive to dynamical solvation effects than ¹⁷O.

Both ¹H and ¹⁷O shieldings undergo a dramatic change from FSS to SSS. Whereas the FSS shieldings are governed by the

paramagnetic hyperfine effects, their magnitude decreases in SSS to a level comparable to the orbital shielding. The second shell deviates, however, still significantly from diamagnetic bulk water in the large σ_{con} contributions of 0.4 and -86 ppm for ¹H and ¹⁷O, respectively.

There is a significant difference between the isotropic dipolar shielding in static calculations of Ni(H₂O)₆²⁺ (Figures 8 and 9) and the average of snapshot calculations. The absolute value of

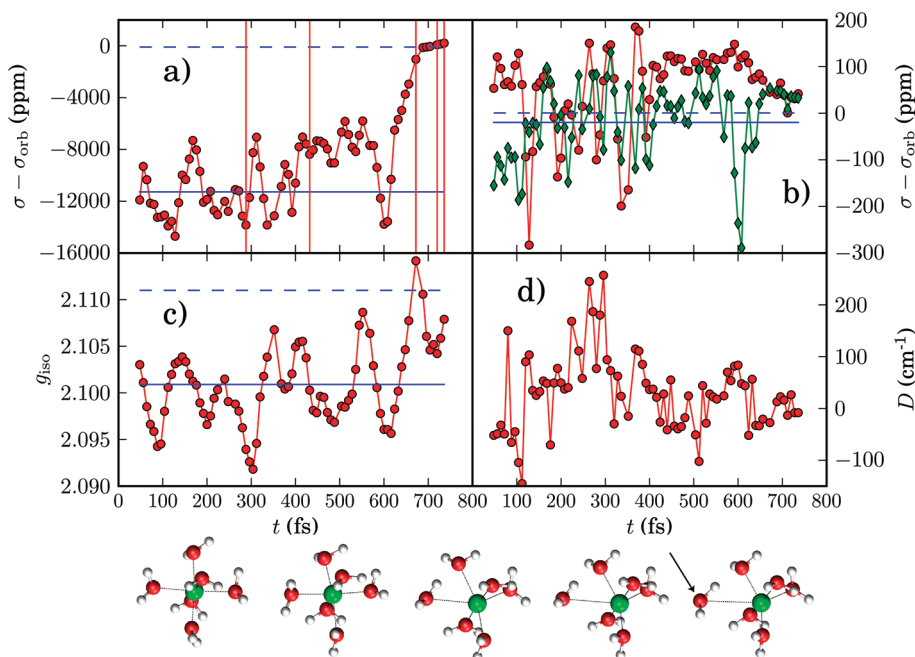


Figure 10. Calculated hyperfine part of the (a) ^{17}O and (b) ^1H isotropic paramagnetic shielding constants (in ppm) together with (c) g value and (d) D parameter of zero-field splitting along the final stages of a first-principles molecular dynamics trajectory of an aqueous solution of the Ni^{2+} ion. The PBE functional was used for a system with 127 water molecules. In particular, the end of the trajectory segment includes the event of the water molecules leaving the first solvation shell. The shielding constants in panels a and b correspond to this molecule (the ^1H shielding constants are plotted for both water protons separately). The five vertical bars in panel a indicate the times appropriate for the depicted structures. The solid blue horizontal lines indicate the first solvation shell averages of the shielding constants, whereas the dashed horizontal lines show values corresponding to the second shell. The arrow at the last depicted structure identifies the departing water molecule. In panel c, the isotropic g value is illustrated together with the averages in 6-fold (solid line) and 5-fold (dashed line) coordinated parts of the trajectory. Panel d visualizes the D parameter of the zero-field splitting tensor.

σ_{dip} is much larger in the latter although still very small compared to the total shielding. This difference is caused by the fluctuation in ZFS in the instantaneous snapshot calculations.

To facilitate comparison with published experimental chemical shift data, we calculated the diamagnetic reference shielding of the H_2O molecule *in vacuo* using the equilibrium structure obtained at the same level and using the same program and settings as were used for the FPMD simulation. The resulting geometrical parameters are $r_{\text{OH}} = 0.978 \text{ \AA}$ and $\angle_{\text{HOH}} = 102.45^\circ$. The shielding calculation using this geometry at the PBE/IGLO-II level yields $\sigma_{\text{O}} = 324.8$ and $\sigma_{\text{H}} = 31.59$ ppm. Furthermore, we used the calculated B3LYP values for the gas-to-liquid transition⁶ $\delta_{\text{g} \rightarrow \text{l}}(^1\text{H}) = -5.27$ ppm and $\delta_{\text{g} \rightarrow \text{l}}(^{17}\text{O}) = -41.2$ ppm to obtain a gross estimate of the liquid-state shielding constants of the bulk diamagnetic water, $\sigma_{\text{O}}^{\text{ref}}(1) = 283.6$ ppm, $\sigma_{\text{H}}^{\text{ref}}(1) = 26.32$ ppm. These reference values were used to obtain the shielding value from the published chemical shift data for the aqueous solution of Ni^{2+} ($\sigma_{\text{exp}} = \sigma_{\text{ref}} - \delta_{\text{exp}}$). To obtain the experimental shielding values at 350 K, at which the presented simulated shielding values were calculated, the experimental shieldings were extrapolated using the $1/T$ dependence of the paramagnetic shielding. The simulated shielding constants in FSS can then be directly compared with the experimental shielding constants σ_{exp} obtained from the chemical shift of the water molecules. The results listed in Table 2 indicate that in the presence of the Ni^{2+} ion the calculated ^{17}O shielding constant agrees with the experimental data of Neely and Connick³⁵ and to a somewhat lesser extent also with those of Fiat and Chmelnick,³⁸ whereas the simulated ^1H shielding constant does not reproduce the experimental data of Granot.³¹

Table 5 lists the simulated contributions of the various physical mechanisms of pNMR shielding to the principal values σ_{ii} ($i = 1, 2, 3$) of the shielding tensor. There are important contributions to the principal values beyond those that are significant for the isotropic values. Despite their negligible isotropic averages, the dipole terms assume the same role here as played by the contact terms in σ_{iso} . The σ_{dip} term contributes to the diagonal values almost comparably to the largest (σ_{con}) term for ^{17}O in FSS. Also, the terms originating in the SO corrections to A and \mathbf{g} , $\sigma_{\text{dip},2}$ and $\sigma_{\text{dip},3}$, respectively, are important for ^{17}O (FSS). For ^{17}O (SSS), the σ_{ii}^{dip} even exceeds σ_{con} in magnitude. The ^1H nuclei in SSS obtain shielding eigenvalues of σ_{dip} larger by 1 or 2 orders of magnitude than the corresponding contact terms. In the present isotropic solution, these large anisotropic dipolar terms with negligible isotropic averages only contribute to NMR relaxation, and not to the structure of static NMR spectra. The water molecules in SSS still experience significantly anisotropic shielding interactions originating from the paramagnetic ion. We note that despite the fact that σ_{con} is formally anisotropic in the novel nuclear shielding theory,²⁵ in practice its eigenvalues are equal within the present statistical error margins. The σ_{ac} term, which is formally not contributing to the isotropic σ , is negligible also for the eigenvalues. [At this point, it is worth it to point out the effect of the averaging described in Table 5. When the principal axis system (PAS) of the individual shielding contribution is not identical to the PAS of the total shielding, the reported projected average values tend to be biased toward their respective isotropic values. After analyzing this issue, we conclude that the majority of the shielding contributions have closely similar PASs on average. Among paramagnetic contributions, there are two exceptions,

$\sigma_{c,\text{aniso}}$ and $\sigma_{p,c}$. Both are related to the g -shift tensor anisotropy ($\Delta\vec{g}$). Those yield in their own PAS average principal values roughly 1 order of magnitude larger compared to those numbers present in Table 5. Since even these numbers are small compared to the principal numbers of the total shielding, we do not report a more detailed analysis here. The PAS of the “diamagnetic” orbital shielding is expectedly also different from the total shielding, and the effect is different in FSS and SSS, according to the relative magnitude of the orbital and paramagnetic shielding in these shells.] The ^1H orbital shielding eigenvalues reveal that despite similar isotropic averages in FSS, SSS, and diamagnetic water, the tensors in FSS and SSS differ significantly.

Dipolar Part of the Hyperfine Interaction. Due to the fundamental importance of A_{dip} in paramagnetic relaxation studies, we present here a few notes about its calculation in the simple point-dipole approximation (PDA) as opposed to the present QM method. Correlation graphs between the QM (PBE) calculation and PDA are plotted in Figure S2, Supporting Information. It can be seen that for ^1H , PDA becomes a very good approximation for the smallest values of A_{33}^{dip} (≈ 1.3 MHz) in the most distant regions of the SSS (≥ 5 Å). On the other hand, the PDA for ^{17}O cannot be considered realistic. [We can anticipate that the quantum-mechanically obtained spin density at the ^{17}O nucleus is qualitatively correct from the agreement between the calculated and experimental contact contributions.] A similar conclusion can be obtained from Table S10, Supporting Information, where the angle between the main principal axis of A_{dip} for ^1H and ^{17}O and corresponding to the Ni–H or Ni–O axes is evaluated. In all of these cases, the two axes are on average close to being parallel. In SSS, however, this applies only for ^1H and the distribution of the angle is sufficiently narrow so that the geometry-based PDA may be useful. It is noteworthy that the O–H bond direction has no apparent significance for the main principal axis of $A_{\text{dip}}(^1\text{H})$.

The evolution along the MD trajectory of the largest principal value A_{33}^{dip} is plotted in Figure S3, Supporting Information. For each case, only one atom was selected. The changes at the end of the traces correspond to the transition from the 6- to the 5-fold-coordinated complex. In the case of ^1H SSS, the evolution of A_{33}^{dip} could be visually related to the evolution of the distance from the Ni^{2+} ion.

Magnetic Properties during the Exchange Event. The hyperfine shielding for the molecules departing from FSS as well as the evolution of the g value and the D parameter of ZFS are followed along the last part of the FPMD trajectory. Figure 10 shows the magnetic parameters at selected snapshots during the departure of the water molecule from FSS. It is seen that the shielding values approach the averages of SSS. In particular, the plot reflects the oscillations of the isotropic part of the hyperfine shielding. The corresponding fluctuations in the full shielding tensor are responsible for Curie-type pNMR relaxation. Likewise, Figure 10c illustrates the step taking place in the g value from the 6-fold to the 5-fold-coordinated situation.

CONCLUSIONS

Using first-principles computations, we have investigated the magnetic properties of the aqueous solution of Ni^{2+} . Both static quantum-chemical calculations of the hexa- and penta-aqua complexes $\text{Ni}(\text{H}_2\text{O})_n^{2+}$ ($n = 5, 6$), as well as first-principles molecular dynamics simulation of the aqueous solution, were performed to obtain structures for the calculation of magnetic

properties. These include all factors affecting the paramagnetic nuclear shielding tensor. The first principles trajectory used for snapshot calculations is the longest of its kind so far, for systems and properties of the present type.

We tested several combinations of methods for the calculation of ESR properties (hyperfine couplings, zero-field splitting, and g tensor), which were subsequently used for the calculation of paramagnetic NMR shielding according to the recent general theory. The spin density distribution of the solvated Ni^{2+} decays rapidly to zero over the first and second solvation shells. We observe that the dominant Fermi contact hyperfine interaction of the hydrogen atoms in the first solvation shell is very sensitive to structure, in particular the tilt angle of the water molecules. Depending on the method used for the static structures, results in the range covering 1 order of magnitude, depending on the tilt angle, are observed. It therefore proved to be beneficial to combine methods for structure calculation and calculation of properties such that both share the same equilibrium geometry, to reduce the range of results and to obtain a well-defined ^1H isotropic hyperfine data. To the contrary, the dipolar ^1H as well as both the contact and dipolar ^{17}O results are less sensitive to the methods used. In particular, the static calculations indicated little influence of the implicit solvation model at the property evaluation step, and the primary influence arises indirectly via the structure.

We compared properties calculated using the 5- and 6-fold-coordinated model structures as well as 5- and 6-fold-coordinated parts of the FPMD trajectory. Upon the transition from 6- to 5-fold-coordination, the eigenvalues of $A_{\text{dip}}(^1\text{H})$ experience a significant increase, whereas the change in $A_{\text{con}}(^1\text{H})$ is very small. The $A_{\text{dip}}(^{17}\text{O})$ eigenvalues are slightly decreased upon entering the 5-fold-coordinated structures, whereas the $A_{\text{con}}(^{17}\text{O})$ values decrease more significantly. The same behavior is followed by the corresponding shielding constants. The significant change of the isotropic ^{17}O shielding constant between the 6- and 5-fold-coordinated Ni^{2+} may be of importance from the experimental point of view. For example, high-pressure NMR determination of the parameters of water exchange in the aqueous Ni^{2+} complex assume σ_{iso} (as well as the underlying A_{iso} and g_{iso}) to be invariant for the different coordination cases. For A_{dip} , we also show that the point-like approximation becomes a good approximation for ^1H in the distant parts of the second solvation shell, at distances around 5 Å and further.

We have followed the g and ZFS tensors along the MD trajectory. Both parameters are very sensitive to the change of geometry and feature short correlation times. A notable feature at the water exchange event is the increase of the perpendicular components of the g tensor upon the transition to a five-coordinated Ni^{2+} ion. The trajectory averages of the g value are slightly above the values calculated for the optimized static 5- and 6-fold coordinated structures. The magnitude of the D parameter decreases significantly upon the transition to the 5-fold coordination. The same behavior is not seen for the static structure, where all 6-fold coordinated models are highly symmetric and show therefore very small zero-field splitting.

Both ^1H and ^{17}O isotropic shielding constants undergo a large change from the first to the second solvation shell. Whereas the FSS shieldings are dominated by the paramagnetic hyperfine effects, in SSS the orbital shielding is already the larger contribution. The situation is different for the shielding anisotropy, as seen from the principal values of the shielding tensors. Here, the dipolar contribution still is much larger than the orbital shielding

for ^1H in SSS. For ^{17}O , the principal values of σ_{dip} (and σ_{con}) are nearly of the same magnitude as σ_{orb} . Several new isotropic shielding contributions (as compared to the case of doublet³⁹) predicted for a higher-than-doublet system²⁵ proved to be very small for the current triplet system.

■ ASSOCIATED CONTENT

S Supporting Information. An extensive methods section, a brief summary of the theory of pNMR shielding, tables with numerical data, and additional figures. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jiri.mares@oulu.fi

■ ACKNOWLEDGMENT

The authors thank Dr. Perttu Lantto and Dr. Matti Hanni (Oulu) for helpful discussions. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 254552. (J.M.) Further support was received from Swiss National Funds—project PBZHP2-125353 (J.M.), University of Helsinki Research Funds, and the University of Oulu. The authors belong to the Finnish Center of Excellence in Computational Molecular Science (CMS). Computational resources due to CSC (Espoo, Finland) were used.

■ REFERENCES

- (1) Hermansson, K.; Knuts, S.; Lindgren, J. *J. Chem. Phys.* **1991**, *95*, 7486.
- (2) Eggenberger, R.; Gerber, S.; Huber, H.; Searles, D.; Welker, M. *J. Chem. Phys.* **1992**, *97*, 5898.
- (3) Malkin, V. G.; Malkina, O. L.; Steinebrunner, G.; Huber, H. *Chem.—Eur. J.* **1996**, *2*, 452.
- (4) Bühl, M.; Parrinello, M. *Chem.—Eur. J.* **2001**, *7*, 4487.
- (5) Bühl, M.; Mauschick, F. T. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5508.
- (6) Pennanen, T. S.; Vaara, J.; Lantto, P.; Sillanpää, A.; Laasonen, K.; Jokisaari, J. *J. Am. Chem. Soc.* **2004**, *126*, 11093.
- (7) Ramalho, T. C.; Bühl, M. *Magn. Reson. Chem.* **2005**, *43*, 139.
- (8) Vaara, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399.
- (9) Asher, J. R.; Kaupp, M. *Theor. Chem. Acc.* **2008**, *119*, 477.
- (10) Schmidt, J.; Hutter, J.; Spiess, H.; Sebastiani, D. *ChemPhysChem* **2008**, *9*, 2313.
- (11) Odelius, M.; Ribbing, C.; Kowalewski, J. *J. Chem. Phys.* **1995**, *103*, 1800.
- (12) Yazyev, O. V.; Helm, L. *J. Chem. Phys.* **2006**, *125*, 54503.
- (13) Yazyev, O. V.; Helm, L. *J. Chem. Phys.* **2007**, *127*, 84506.
- (14) Lindgren, M.; Laaksonen, A.; Westlund, P. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10368.
- (15) Kowalewski, J.; Larsson, T.; Westlund, P. O. *J. Magn. Reson.* **1987**, *74*, 56.
- (16) Odelius, M.; Ribbing, C.; Kowalewski, J. *J. Chem. Phys.* **1996**, *104*, 3181.
- (17) Kowalewski, J.; Egorov, A.; Kruk, D.; Laaksonen, A.; Aski, S. N.; Parigi, G.; Westlund, P. *J. Magn. Reson.* **2008**, *195*, 103.
- (18) Kruk, D.; Kowalewski, J. *J. Chem. Phys.* **2002**, *116*, 4079.
- (19) Kruk, D.; Kowalewski, J.; Westlund, P. *J. Chem. Phys.* **2004**, *121*, 2215.
- (20) Kruk, D.; Kowalewski, J. *J. Chem. Phys.* **2009**, *130*, 174104.
- (21) Yazyev, O. V.; Helm, L. *Eur. J. Inorg. Chem.* **2008**, *2008*, 201.

- (22) Clore, G. M.; Iwahara, J. *Chem. Rev.* **2009**, *109*, 4108.
- (23) Mareš, J.; Liimatainen, H.; Laasonen, K.; Vaara, J. *J. Chem. Theory Comput.* **2011**, in press; doi: 10.1021/ct200320z.
- (24) Kowalewski, J.; Mäler, L. *Nuclear Spin Relaxation in Liquids: Theory, Experiments, and Applications*, 1st ed.; Taylor & Francis: United Kingdom, 2006; p 440.
- (25) Pennanen, T. O.; Vaara, J. *Phys. Rev. Lett.* **2008**, *100*, 133002.
- (26) Munzarová, M.; Kaupp, M. *J. Phys. Chem. A* **1999**, *103*, 9966.
- (27) Neese, F. ORCA; University of Bonn: Bonn, Germany, 2009.
- (28) Frisch, M. J. et al. *Gaussian 03*, revision C.02.; Gaussian Inc.: Wallingford, CT, 2009.
- (29) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103.
- (30) Helgaker, T.; Jaszuński, M. *J. Chem. Theory Comput.* **2007**, *3*, 86.
- (31) Granot, J. *J. Chem. Phys.* **1974**, *61*, 3043.
- (32) Gauss, J.; Kállay, M.; Neese, F. *J. Phys. Chem. A* **2009**, *113*, 11541.
- (33) Griffiths, J. H. E.; Owen, J. *Proc. R. Soc. London, Ser. A* **1952**, *213*, 459.
- (34) Friedman, H. L.; Holz, M.; Hertz, H. G. *J. Chem. Phys.* **1979**, *70*, 3369.
- (35) Neely, J. W.; Connick, R. E. *J. Am. Chem. Soc.* **1972**, *94*, 3419.
- (36) Blumberg, W. E. In *Magnetic Resonance in Biological Systems*, 1st ed.; Ehrenberg, A., Malmström, B., Eds.; Pergamon Press: Oxford, U.K., 1967; p 110.
- (37) Neese, F. In *Calculation of NMR and EPR Parameters: Theory and Applications*, 1st ed.; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: New York, 2004; p 541.
- (38) Fiat, D.; Chmelnick, A. M. *J. Am. Chem. Soc.* **1971**, *93*, 2875.
- (39) Pennanen, T. O.; Vaara, J. *J. Chem. Phys.* **2005**, *123*, 174102.

Density Functional Restricted–Unrestricted/Molecular Mechanics Theory for Hyperfine Coupling Constants of Molecules in Solution

Zilvinas Rinkevicius,^{*,†,||} N. Arul Murugan,[†] Jacob Kongsted,[‡] Bogdan Frecuș,[†] Arnfinn Hykkerud Steindal,[§] and Hans Ågren[†]

[†]Department of Theoretical Chemistry & Biology, School of Biotechnology, Royal Institute of Technology, SE-106 91 Stockholm, Sweden

[‡]Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

[§]Centre of Theoretical and Computational Chemistry, Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway

^{||}Swedish e-Science Research Center (SeRC), Royal Institute of Technology, SE-100 44 Stockholm, Sweden

ABSTRACT: A density functional restricted–unrestricted approach, capable of evaluating hyperfine coupling constants with the inclusion of spin polarization effects in a spin-restricted Kohn–Sham method, has been extended to incorporate environmental effects. This is accomplished by means of a hybrid quantum mechanics/molecular mechanics formalism which allows for a granular representation of the polarization and electrostatic interactions with the classically described medium. By this technique, it is possible to trace the physical origin of hyperfine coupling constants in terms of spin polarization and spin density contributions and disentangle the dependence of these contributions on molecular geometry and solvent environment, something that increases the prospects for optimal design of spin labels for particular applications. A demonstration is given for the nitrogen isotropic hyperfine coupling constant in di-*tert*-butyl nitroxide solvated in water. The results indicate that the direct spin density contribution is about 5 times smaller than the spin polarization contribution to the nitrogen isotropic hyperfine coupling constant and that the latter contribution is solely responsible for the solvent shift of the constant. The developed approach is found capable of achieving satisfactory accuracy in prediction of the hyperfine coupling constants of solvated di-*tert*-butyl nitroxide and other similar nitroxides without the inclusion of solvent molecules in the quantum region provided polarizable force fields are used for the description of these molecules.

1. INTRODUCTION

Modeling of electron paramagnetic resonance (EPR) spin Hamiltonian parameters of nitroxide spin labels and related compounds has attracted considerable interest since the development of the first electronic structure methods for open shell molecules.^{1–19} Over the years, numerous investigations of electronic \mathbf{g} tensors^{1,2,4,7–11,15,19} and hyperfine coupling tensors^{1–3,5,6,8–18} have been carried out, and the mechanisms governing the behavior of EPR parameters of radicals in a vacuum and solutions have been thoroughly studied. Despite these achievements, the accurate prediction of electronic \mathbf{g} tensors and especially hyperfine coupling constants of nitrogen in various nitroxides in protic solvents, like water or methanol, remains a challenge, as it requires a simultaneous account of solute vibrational degrees of freedom and solvent dynamics.^{8,11–13,15,16,19} Recently, Barone and co-workers suggested a so-called integrated approach to overcome this kind of difficulty in the modeling of molecular properties²⁰ and successfully applied it to study several prototypical nitroxide radicals.^{8,10,11,15,16} The essence of Barone et al.'s integrated approach is that the solute/solvent dynamics at a selected temperature are simulated using classical or Car–Parrinello (CP) molecular dynamics (MD) with the spin Hamiltonian parameters evaluated by means of density functional theory (DFT) over a set of snapshots extracted from the MD trajectory. Nowadays, approaches of this kind have been adopted by several

groups interested in EPR spin Hamiltonian parameters of solvated radicals,^{12,13,19,21,22} generating a steady increase of investigations.

The first part of the computations using the integrated approach, namely, the simulation of solute/solvent dynamics, is typically carried out by means of Car–Parrinello MD for short time scales, in the range of picoseconds,^{8,11,19} or by means of classical MD for longer time scales, in the range of nanoseconds.^{12–14} However, in the latter case, the applications are restricted to radicals for which classical force fields have been developed.^{23,24} The second part of the computations in the integrated approach—the evaluation of electronic \mathbf{g} tensors and hyperfine coupling tensors—has mostly been accomplished using unrestricted DFT combined with the polarizable continuum model (PCM) to account for the solvent environment.^{8,11} Recently, hybrid density functional theory/molecular mechanics (DFT/MM) has emerged as an alternative to unrestricted DFT/PCM, and several applications of such methods for EPR spin Hamiltonian parameters of solvated nitroxides have now appeared.^{12–14,19} Previously, one of us published an advanced QM/MM method based on the polarizable embedding scheme,²⁵ and in this paper, we extend this method to include calculations of hyperfine coupling constants of radicals.

Received: May 28, 2011

Published: August 17, 2011

Following our previous works devoted to the computation of hyperfine coupling constants of organic radicals and transition metal complexes,^{26–28} we employ a density functional restricted–unrestricted (DFT-RU) approach and extend it into a hybrid quantum mechanics/molecular mechanics (QM/MM) setting. This work complements a recently implemented DFT/MM approach for the evaluation of electronic \mathbf{g} tensors¹⁹ and thus allows one to study the complete set of EPR spin Hamiltonian parameters of solvated nitroxide spin labels and other radicals with doublet ground states. Furthermore, this opens the possibility to substitute the conventional DFT/PCM method with the hybrid DFT/MM method in the integrated approach for studies of EPR spin Hamiltonian parameters in various environments, like proteins and cellular membranes, and thus to study such environmental effects on these parameters in greater detail and at a higher accuracy.

2. THE DENSITY FUNCTIONAL RESTRICTED–UNRESTRICTED APPROACH FOR SOLVENT ENVIRONMENTS

Spin-restricted DFT is designed to be spin-contamination-free, thus providing a starting formalism for EPR spin Hamiltonian parameters that from a conceptual point of view is more appropriate than the conventionally used unrestricted DFT, as it allows one to obtain a strict one-to-one mapping between the electron density and spin state of the molecule.^{26–28} Despite this advantage, spin-restricted DFT methods have seldomly been used for the computation of hyperfine coupling constants, since they are incapable of accounting for spin polarization effects, which are of crucial importance for accurate prediction of these constants. Recently, we developed an extension of spin-restricted DFT—the so-called restricted–unrestricted approach, which follows the principles outlined in pioneering works by Fernández et al.^{29,30} devoted to spin polarization effects in multiconfigurational self-consistent field calculations of molecular properties and allows one to account for spin polarization effects in calculations of linear and nonlinear molecular properties within a spin-restricted Kohn–Sham formalism.^{26–28} This approach has been successfully applied to study electronic \mathbf{g} tensors as well as hyperfine coupling constants of various organic radicals and transition metal complexes and has arguably provided comparable or better accuracy than the widely used unrestricted DFT methods. Our development of the DFT-RU approach has been focused on benchmarking its performance, and the modeling has thus been limited to a cluster approach in which the solvent molecules are treated explicitly along with the solute. In order to provide a more refined tool for investigation of the environmental effects on hyperfine coupling constants within the DFT-RU formalism, we implement in this work the hybrid DFT-RU/MM formalism based on the DFT/MM approach developed by Olsen et al.²⁵ In the remaining part of this section, we briefly review the basic principles behind the DFT-RU approach and describe the DFT-RU/MM extension, now implemented in a development version of the DALTON program.³¹

In DFT-RU, electronic spin-dependent properties, like hyperfine coupling constants, are typically computed as the sum of two terms:^{26–28} one which accounts for the direct spin density contribution and the second which accounts for the spin polarization contribution to the property. More specifically, the expectation value in this approach for an arbitrary one-electron

spin-dependent operator \hat{A} can be written as

$$\langle \hat{A} \rangle = \langle \hat{A} \rangle_{\text{den}} + \langle \hat{A} \rangle_{\text{pol}} \\ = \frac{\partial E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]}{\partial x} + \tilde{\lambda} \frac{\partial^2 E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]}{\partial x \partial \tilde{\lambda}} \quad (1)$$

where $\langle \hat{A} \rangle_{\text{den}}$ and $\langle \hat{A} \rangle_{\text{pol}}$ are the direct spin density and spin polarization contributions to the expectation value of operator \hat{A} , and $E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]$ is the energy functional of the molecular system in the presence of a perturbation described by operator \hat{A} , which is dependent on the perturbed electron density $\rho(\mathbf{r}, x)$, perturbation strength x , and set of Lagrangian multipliers $\tilde{\lambda} = (\bar{\kappa}_{pq}, \bar{t}_{pq})$ associated with singlet and triplet orbital rotations, which are collected into the set $\lambda = (\kappa_{pq}, t_{pq})$. The computation of $\langle \hat{A} \rangle_{\text{den}}$ is a trivial task and is accomplished by contracting the unperturbed density $\rho(\mathbf{r}, x = 0)$ with operator \hat{A} , while the evaluation of $\langle \hat{A} \rangle_{\text{pol}}$ presents a more involved undertaking, as it requires the determination of Lagrangian multipliers $\tilde{\lambda}$ by solving

$$\frac{\partial^2 E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]}{\partial \tilde{\lambda}^2} \tilde{\lambda} = - \left. \frac{\partial E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]}{\partial \tilde{\lambda}} \right|_{x=0} \quad (2)$$

Generally, in the approximate DFT-RU approach, the coupling between singlet and triplet orbital rotations can be neglected without a substantial loss of accuracy,²⁷ and thus the problem of finding Lagrangian multipliers $\tilde{\lambda}$ reduces to the task of finding Lagrangian multipliers associated with triplet orbital rotations by solving a set of linear equations

$$\frac{\partial^2 E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]}{\partial \tilde{t}_{pq}^2} \tilde{t}_{pq} = - \left. \frac{\partial E[\rho(\mathbf{r}, x), \tilde{\lambda}, x]}{\partial \tilde{t}_{pq}} \right|_{x=0} = -\tau_{pq} \quad (3)$$

where the energy functional gradient τ_{pq} with respect to triplet orbital rotations t_{pq} is nonvanishing in a spin-restricted Kohn–Sham formalism. The procedures for evaluation of triplet gradient elements, τ_{pq} , and solving eq 2 or eq 3 are described in detail in our previous works devoted to the DFT-RU approach.^{26–28} In the following, we outline modifications of the DFT-RU approach in the spirit of the recently developed DFT/MM method of Olsen et al.²⁵ required for enabling investigations of hyperfine coupling constants of solvated radicals.

The starting point to derive the DFT-RU/MM method is the Kohn–Sham Hamiltonian used by Olsen et al.²⁵ consisting of the conventional Kohn–Sham Hamiltonian describing the quantum region and a special coupling term between the quantum (QM) and classical (MM) regions

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{QM/MM}} \quad (4)$$

where \hat{H}_{QM} is defined in second quantization notation as

$$\hat{H}_{\text{QM}} = \sum_{pq} f_{pq} \hat{E}_{pq} = \sum_{pq} (h_{pq} + j_{pq} + v_{pq}^{\text{xc}}) \hat{E}_{pq} \quad (5)$$

and where the coupling term $\hat{H}_{\text{QM/MM}}$, which includes electrostatic and polarization interactions between QM and MM regions, is defined according to Olsen et al.²⁵ as

$$\hat{H}_{\text{QM/MM}} = \sum_{pq} \sum_s M_{s,pq}^{(n)} \hat{E}_{pq} - \sum_{pq} \sum_a \mu_a^{\text{ind}} \mathbf{t}_{pq}^a \hat{E}_{pq} \quad (6)$$

In the above equations, we use notations frequently employed in density functional response theory for various matrix elements: f_{pq}

is the matrix element of the Kohn–Sham operator, which consists of the “core” Hamiltonian term h_{pq} i.e., the sum of one-electron kinetic energy and nuclear attraction operators matrix elements, the Coulomb interaction term j_{pq} and the exchange–correlation term v_{pq}^{xc} . $M_{s,pq}^{(n)}$ is an electrostatic interaction operator matrix element due to the n th order multipole at the site s in the MM region (see eq 8 in ref 25 for the specific expression); μ_a^{ind} is the induced dipole moment at the polarizable MM site a created by the total (QM and MM) electric field at this site; t_{pq}^a is the matrix element of the electric field operator for site a in the MM region; $\hat{E}_{pq} = p_{\alpha}^{\dagger}q_{\alpha} + p_{\beta}^{\dagger}q_{\beta}$ is the singlet excitation operator, which is in second quantization written as the combination of the spin-dependent creation operators for orbital p and spin-dependent annihilation operators for orbital q . For a more detailed description of the enumerated quantities as well as the optimization procedures for the QM density and the MM induced dipole moments, we refer to the original work on the DFT/MM formalism.²⁵

After describing the QM and QM/MM terms in the Kohn–Sham Hamiltonian of the unperturbed system, let us consider the situation in which the QM region also is affected by an electronic spin-dependent perturbation like the electron–nucleus Fermi contact interaction operator. In this case, the Hamiltonian becomes

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{QM/MM}} + x\hat{A} \quad (7)$$

where the spin-dependent perturbation operator in the general form can be written as

$$\hat{A} = \sum_m (-1)^m \sum_{pq} V_{pq}^{-m} T_{pq}^m \quad (m = -1, 0, 1) \quad (8)$$

with V_{pq}^{-m} being the spherical component of the matrix element and T_{pq}^m being the component of the triplet excitation operator associated with a corresponding element of electronic spin. Taking the form of the perturbation operator, the electron density of the perturbed system can, as shown in our previous works on the DFT-RU method,^{26–28} be parametrized as

$$|\tilde{0}\rangle = \sum_{p > q} \exp[-\kappa_{pq}(\hat{E}_{pq} - \hat{E}_{qp})] \exp[-t_{pq}(\hat{T}_{pq} - \hat{T}_{qp})] |0\rangle \quad (9)$$

where κ_{pq} and t_{pq} are the parameters associated with singlet and triplet orbital rotations and T_{pq} is the triplet excitation operator, i.e., $\hat{T}_{pq} = p_{\alpha}^{\dagger}q_{\alpha} - p_{\beta}^{\dagger}q_{\beta}$. Adopting this parametrization of the Kohn–Sham determinant in the case of the DFT-RU/MM approach, we can rewrite eq 3, which defines the spin polarization correction to the expectation value of the spin-dependent operator, as the set of linear equations

$$\mathbf{E}^{[2]}\bar{\mathbf{r}} = \langle\langle 0 | [\hat{\mathbf{q}}, [\hat{\mathbf{q}}^{\dagger}, \hat{H}_{\text{QM}}] + \hat{H}_{\text{QM}}^{(1)}] | 0 \rangle\rangle + \langle\langle 0 | [\hat{\mathbf{q}}, [\hat{\mathbf{q}}^{\dagger}, \hat{H}_{\text{QM/MM}}] + \hat{H}_{\text{QM/MM}}^{(1)}] | 0 \rangle\rangle \bar{\mathbf{r}} = -\tau \quad (10)$$

where $\hat{H}_{\text{QM}}^{(1)}$ and $\hat{H}_{\text{QM/MM}}^{(1)}$ are the QM and QM/MM contributions to the first-order perturbed Hamiltonian, the column $\bar{\mathbf{r}}$ is the collection of Lagrangian multipliers \bar{r}_{pq} for triplet orbital rotations, $\bar{\mathbf{q}}^{\dagger}$ is the column vector of triplet excitation operators \hat{T}_{pq} , and τ is the energy functional $E[\rho(\mathbf{r}, 0), \bar{\lambda}, 0]$ with respect to

triplet orbital rotations with elements

$$\tau_{pq} = 2\langle 0 | [\hat{T}_{pq}, \hat{H}_{\text{QM}}] | 0 \rangle + 2\langle 0 | [\hat{T}_{pq}, \hat{H}_{\text{QM/MM}}] | 0 \rangle \quad (11)$$

In the latter two equations, we separated the orbital Hessian matrix $\mathbf{E}^{[2]}$ and triplet gradient τ_{pq} into QM and QM/MM contributions according to the partitioning of the Kohn–Sham Hamiltonian in eq 4. Procedures for determination of the triplet Lagrangian multipliers are described in detail in our previous works devoted to the DFT-RU approach.^{26–28} The only two new elements introduced in eq 10 by extending the DFT-RU approach to a QM/MM setting are the QM and MM region coupling contributions from the Hamiltonian $\hat{H}_{\text{QM/MM}}$ to the $\mathbf{E}^{[2]}$ matrix and the τ column vector. The procedure for evaluation of the QM/MM contribution to the orbital Hessian in the DFT-RU/MM approach is exactly the same as the one used for linear triplet response functions in the DFT/MM formalism and has already been implemented in the DALTON program in connection with DFT/MM methods for the evaluation of nuclear spin–spin coupling constants³² and triplet excitations energies.³³ Thus, only the $\hat{H}_{\text{QM/MM}}$ contribution to the triplet gradient (see eq 11) has been left to be implemented in this work in order to extend DFT-RU to the DFT-RU/MM setting. We accomplished this task by adapting subroutines used for evaluation of the orbital Hessian for the triplet response function to computation of the triplet gradient of the form given in eq 11.

After discussing the electronic spin-dependent properties within spin-restricted density functional theory using the restricted–unrestricted approach and presenting the details of extension of this formalism to the QM/MM setting, we describe in the remaining part of this work an application of the DFT-RU/MM method to study the isotropic hyperfine coupling constants of nitrogen in di-*tert*-butyl nitroxide (DTBNO) in aqueous solution.

3. COMPUTATIONAL DETAILS

For assessment of the DFT-RU/MM approach in prediction of the hyperfine coupling constants (HFCCs) of solvated radicals, we opted to study the isotropic HFCC of nitrogen in the prototypical nitroxide, namely di-*tert*-butyl nitroxide, in an aqueous environment. This choice is motivated by the availability of recent and detailed studies of nitrogen HFCCs in this system using unrestricted DFT/PCM and DFT/MM methods^{8,13} as well as by results of our investigation of the electronic \mathbf{g} tensor of this system using the DFT/MM formalism.¹⁹ The latter work showed exceptional capabilities of the DFT/MM method to model a solvent environment using polarizable force fields. In this work, we determined the isotropic HFCC of nitrogen in DTBNO in aqueous solution at ambient temperature using the integrated approach proposed by Barone et al.²⁰ Thus, as the first step, a molecular dynamics simulation of DTBNO in water has been carried out using the hybrid Car–Parrinello/molecular mechanics method, and as the second step, the isotropic HFCC of nitrogen has been computed using the DFT-RU/MM method over a set of snapshots extracted from the MD trajectory generated in the first step. Details of both steps of computations for di-*tert*-butyl nitroxide in aqueous solution are given below.

3.1. Hybrid Car–Parrinello/Molecular Mechanics Simulation of Di-*tert*-butyl Nitroxide in Aqueous Solution. We have employed the hybrid Car–Parrinello/molecular mechanics (CP-MD/MM) approach^{34,35} to generate the structures of di-*tert*-butyl nitroxide in aqueous solution at the ambient temperature. The details of this CP-MD/MM simulation have already been

given in our previous work¹⁹ devoted to electronic \mathbf{g} tensors of DTBNO, but for the sake of completeness, we repeat some key points here. In the CP-MD/MM simulation, only DTBNO has been treated at the density functional theory level using the BLYP exchange-correlation functional^{36,37} and Troullier–Martins norm-conserving pseudopotentials³⁸ for the core electrons, and all 9679 waters have been moved to the MM region and described by the TIP3P force field.³⁹ The solute and solvent system, i.e., “DTBNO+9679 waters” has been included in an orthorhombic box with a size of approximately $68.4 \text{ \AA} \times 65.8 \text{ \AA} \times 65.4 \text{ \AA}$, and the production run has been carried out for 21 ps in the canonical ensemble, where the temperature has been controlled using a Nosé–Hoover thermostat.^{40,41} During this simulation, the following parameters were employed for the treatment of the QM region: the fictitious electronic mass was set to 600 amu; the integration time step was selected to be 5 au; the electronic wave function expansion cut off was set to 80 Ry. The initial structure for starting CP-MD/MM simulation was obtained from classical MD simulations as described in our previous work devoted to solvent effects on the electronic \mathbf{g} tensor of DTBNO.¹⁹

3.2. DFT-RU/MM Computations of Isotropic Hyperfine Coupling Constant of Nitrogen in Di-*tert*-butyl Nitroxide.

According to the philosophy of the integrated approach, we computed nitrogen (¹⁴N isotope) isotropic hyperfine coupling constant, a_N , in DTBNO solvated in water by averaging a_N evaluated using DFT-RU/MM over 86 snapshots taken from the CP-MD/MM trajectory. In these calculations and in analogy to our previous work devoted to the electronic \mathbf{g} tensor of DTBNO,¹⁹ we exploited two QM region models: one in which only DTBNO is included in the QM region and another in which DTBNO and two water molecules closest to the NO bond are included in the QM region. Similarly, the MM region has been designed to include all water molecules within the 20 Å radius of the DTBNO in order to ensure a converged description of the solvent environment, and five different types of force fields have been used to describe water molecules in the MM region. The selected force fields, denoted as MM- X ($X = 0-4$) in Table 1, follow the force fields ladder established in our previous work;¹⁹ going from the MM-0 to the MM-4 force field provides a more accurate description of the water potential. Therefore, we carried out five different sets of calculations of the nitrogen HFCC for each QM region model by changing the force field description of the MM region. These calculations allow us to investigate the HFCC dependence on the accuracy of the MM description and to establish a requirement for the quality of the force field in the DFT-RU/MM calculations.

After giving the details of the MM region used in our DFT-RU/MM calculations, we turn to the QM region. In all calculations, we used the hybrid B3LYP^{36,37,42,43} and PBE0^{44,45} exchange–correlation functionals, which are known to be some of the most accurate functionals for evaluation of HFCCs in various organic radicals.^{17,46–48} The second crucial choice is the selection of the basis set, as this basis set should be capable of describing accurately the electron density in the vicinity of the nucleus and at the same time be flexible enough in the outer valence region to account for spin polarization effects. One way to fulfill these requirements is to select a basis set augmented with tight s -type functions and several sets of polarization functions. We successfully employed this strategy in our previous studies of HFCCs of various organic radicals and paramagnetic transition metal complexes.^{26–28} However, none of these studies included the nitroxides, and thus convergence of the nitrogen HFCCs

Table 1. Hierarchy of the Force Fields Used to Describe the Water Molecules in DFT-RU/MM Calculations of Nitrogen Hyperfine Coupling Constant in Di-*tert*-butyl Nitroxide in Aqueous Solution

force field	electrostatics		polarization		ref
	multipoles	expansion points	polarization	expansion points	
MM-0	charges	oxygen atom hydrogen atoms	none	none	39
MM-1	charges	oxygen atom hydrogen atoms	isotropic	oxygen atom	60
MM-2	charges	oxygen atom hydrogen atoms	anisotropic	oxygen atom hydrogen atoms	19
MM-3	charges dipoles quadrupoles	oxygen atom hydrogen atoms OH midbonds ^a	anisotropic	oxygen atom hydrogen atoms OH midbonds	19
MM-4	charges dipoles quadrupoles octupoles	oxygen atom hydrogen atoms OH midbonds ^a	anisotropic	oxygen atom hydrogen atoms OH midbonds	19

^a Are not used as the expansion points for charges.

with respect to Huz- X ($X = \text{III,IV}$) type basis sets^{49,50} has not been investigated. In order to select a suitable basis set for our DFT-RU/MM calculations, we carried out an investigation of the basis set dependence of a_N in the “DTBNO + 2 waters” system, in which the behavior of both the direct spin density contribution, a_N^{den} , and the spin polarization contribution, a_N^{pol} , with respect to basis set quality has been established. The results of this investigation are tabulated in Table 2 and indicate that a_N converges going from the Huz-III basis set to the Huz-IIIsu3 basis set up to 0.04 G independently on the exchange–correlation functional used in calculations. Here, we point out that already using the Huz-III basis set an accurate description of the direct density contribution to a_N is achieved and that only the spin polarization contribution is affected by improved description of the core region going from the Huz-III to the Huz-IIIsu3 basis set. Upon going from the Huz-III type basis sets to Huz-IV type basis sets, the description of the spin polarization contribution is further improved, indicating the importance of polarization functions for the accurate description of this contribution. The results clearly demonstrate that an accurate description of the spin polarization contribution is achieved only if the basis set used provides a balanced description of the inner core region and the outer valence region of atoms. Taking into account these findings and considering the computational cost, we performed all DFT-RU/MM calculations in this work using the Huz-IIIsu3 basis set. The selection of the exchange-correlation functional and basis set for description of the QM region in our DFT-RU/MM calculations is very similar to the one used in our studies of HFCCs in various organic radicals using the DFT-RU approach^{26–28} and thus according to our experience should be adequate for nitrogen isotropic HFCCs in solvated DTBNO.

4. RESULTS AND DISCUSSION

We analyze in detail the performance of the DFT-RU/MM approach for the nitrogen isotropic hyperfine coupling constant in DTBNO in aqueous solution and compare the obtained

results with available experimental data^{51–53} as well as results of recent studies by Pavone et al.⁸ and by Houriez et al.¹³ To start, we give a brief description of the nitrogen hyperfine coupling constants from the perspective of the DFT-RU approach.

4.1. Isotropic Hyperfine Coupling Constant of Nitrogen in Nitroxides. The EPR spin Hamiltonian parameters in nitroxides are governed to a large extent by the two molecular orbitals localized on the NO moiety of the nitroxide (see Figure 1), namely, a singly occupied orbital (SOMO) of π type in which the unpaired electron resides and a doubly occupied orbital (HOMO) which accommodates a lone pair from oxygen. It has long been recognized that few states models,^{54,55} which involve the SOMO

and HOMO orbitals of nitroxide, explain the experimentally observed ordering of the electronic g -tensor components in these radicals or rationalize the solvent shift of the electronic g tensor in various solvents in terms of a blue shift of the HOMO \rightarrow SOMO excitation and electron density delocalization in the SOMO upon solvation.⁵¹ Furthermore, the size of the nitrogen isotropic HFCC in nitroxides has been associated with the spin density distribution pattern in the SOMO orbital and its solvent behavior interpreted in terms of the selective stabilization of neutral or zwitterionic structure of the ONRR' moiety by solvent molecules. Over the years, this picture of the physical mechanisms remained largely unchanged, and its validity has been verified many times by quantum chemical calculations.^{4,5,7,8,19}

According to the above outlined interpretation, the nitrogen isotropic HFCC, a_N , in nitroxides is mostly dependent on the spin density distribution in the SOMO orbital localized on the NO bond. Thus, a_N is inherently dependent on two geometrical parameters of the ONRR' moiety (see Figure 1): the length of the NO bond and the improper dihedral angle θ , which characterizes the NO bond tilt with respect to the NRR' plane. From these two parameters, only the second one is important for the nitrogen isotropic HFCC, as the variation of the first parameter, the NO bond length, only weakly influences the size of a_N in DTBNO and other nitroxides according to a recent study of Pavone et al.,⁸ while changes in the second parameter, the improper angle θ , strongly affect a_N and increase more than twice with θ going from 0 to $\pm 60^\circ$ (see Figure 2) according to our DFT-RU calculations. Such a sharp dependence of nitrogen isotropic HFCC on the out-of-plane motion of the NO bond has already been observed in previous works,^{8,12} and an accurate description of this effect has been found to be a key factor defining the overall reliability of computed nitrogen HFCCs. However, most of the previous studies have been carried out at the unrestricted DFT level, and thus a detailed analysis of the behavior of the direct spin density and spin polarization contributions to a_N under the NO bond out of plane motion was impossible. In order to fill this gap, we carried out DFT-RU calculations of the nitrogen isotropic HFCC in DTBNO for different values of the improper dihedral angle, $\theta = -60$ to $+60$. In Figure 2, we plotted both direct density, a_N^{den} , and spin

Table 2. Basis Set Dependence of Nitrogen Isotropic HFCC in “DTBNO + 2 waters” System in DFT-RU Calculations^{a,b}

basis set	B3LYP			PBE0		
	a_N^{den} , G	a_N^{pol} , G	a_N , G	a_N^{den} , G	a_N^{pol} , G	a_N , G
Huz-III	2.26	11.92	14.18	2.27	12.32	14.59
Huz-IIIsu ^c	2.23	11.84	14.07	2.24	12.42	14.66
Huz-IIIsu1 ^d	2.26	12.03	14.30	2.27	12.64	14.91
Huz-IIIsu2 ^e	2.26	12.12	14.39	2.29	12.71	15.00
Huz-IIIsu3 ^f	2.28	12.14	14.43	2.29	12.75	15.04
Huz-IV	2.28	12.17	14.45	2.22	12.92	15.14
Huz-IVsu4 ^g	2.27	12.33	14.60	2.28	12.94	15.22

^a Nitrogen isotropic HFCC, a_N , has been computed as the average over 86 snapshots extracted from the CP-MD/MM trajectory. ^b All DFT-RU calculations have been performed using the hybrid B3LYP exchange–correlation functional. ^c All s -type functions have been uncontracted in the Huz-III basis set. ^d All s -type functions have been uncontracted in the Huz-III basis set, and one tight s -type basis function has been added according to procedure described in ref 61. ^e All s -type functions have been uncontracted in the Huz-III basis set, and two tight s -type basis functions have been added according to procedure described in ref 61. ^f All s -type functions have been uncontracted in the Huz-III basis set, and three tight s -type basis functions have been added according to the procedure described in ref 61. ^g All s -type functions have been uncontracted in the Huz-IV basis set, and four tight s -type basis functions have been added according to the procedure described in ref 61.

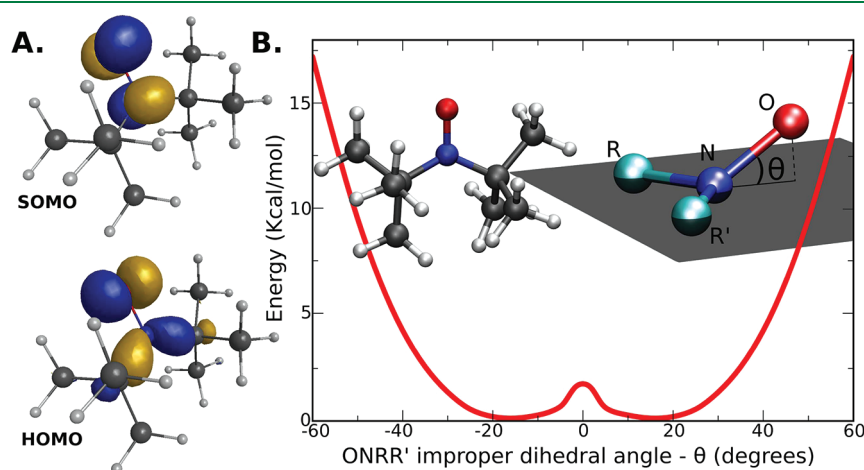


Figure 1. Key features of the electronic and geometrical structure of di-*tert*-butyl nitroxide. (A) Isosurface plots of π -type SOMO and n -type HOMO orbitals localized on the NO bond. (B) Potential energy surface for improper dihedral angle θ obtained by performing a relaxed potential energy surface scan with the spin-restricted open-shell MP2 method⁵⁶ using Ahlrich et al.'s TZVP basis set.⁶² These calculations have been carried out using the GAMESS-US program.⁶³

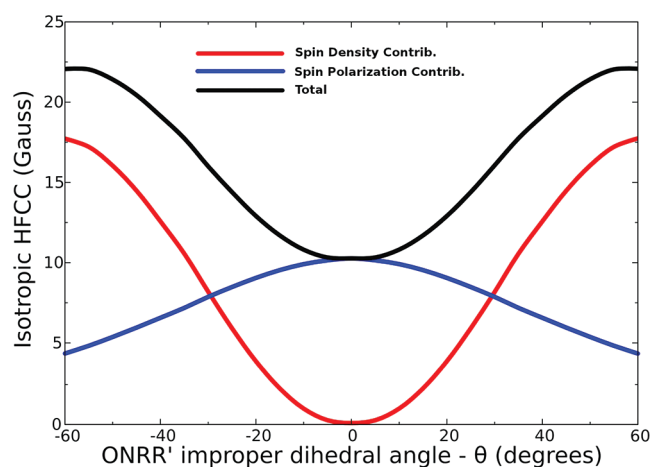


Figure 2. Dependence of nitrogen isotropic HFCC in di-*tert*-butyl nitroxide on the improper dihedral angle θ obtained using DFT-RU method at the B3LYP/Huz-IIIsu3 level.

polarization, a_N^{pol} , contributions to the a_N dependence on improper dihedral angle θ . As expected, Figure 2 shows the typical nitrogen isotropic HFCC dependence on the NO bond out of plane motion; i.e., the a_N dependence on the improper dihedral angle θ has a parabolic shape with a minimum at the planar geometry of the ONRR' moiety, and our DFT-RU results are consistent with previous findings obtained for various nitroxides at the unrestricted DFT level.^{8,12} Here, we would like to point out that, from the two contributions constituting the nitrogen isotropic HFCC, only the direct spin density contribution a_N^{den} exhibits a similar dependence on the improper dihedral angle θ to the isotropic HFCC itself (see Figure 2), while the spin polarization contribution a_N^{pol} behaves completely differently and slowly decreases with the increase of the absolute value of the improper dihedral angle θ (see Figure 2). A closer inspection of Figure 2 reveals that the spin polarization contribution dominates a_N for the small values of the improper dihedral angle θ , as the π -type SOMO orbital localized on the NO bond in such a geometrical arrangement of the ONRR' moiety does not provide efficient pathways for the generation of a sizable direct spin density contribution to a_N . However, for larger values of the improper dihedral angle θ , a_N^{den} sharply increases and becomes larger than the spin polarization contribution a_N^{pol} for $\theta > 30$ (see Figure 2). Thus, the composition of the nitrogen isotropic HFCC in DTBNO as well as other nitroxides is directly controlled by the nonplanarity of the ONRR' moiety, and for typical values of the improper dihedral angle $\theta = 10$ – 20° , the nitrogen isotropic HFCC is dominated by the spin polarization contribution, while the direct spin density contribution plays only a minor role amounting only up to 10–30% of a_N values. The prominent role of spin polarization in nitrogen isotropic HFCC is an important finding, which gives a different interpretation of the shift of a_N upon solvation as an adjustment to the change of spin density relaxation caused by solute interaction with solvent molecules. We will return to this topic and discuss it in the section devoted to the solvent shift of nitrogen isotropic HFCCs.

There are two different views on the shape of the potential energy surface (PES) associated with the NO bond out of plane motion. It is widely accepted that in most nitroxides the ONRR' moiety is nonplanar and that the PES with respect to this degree

of freedom is a symmetric or an asymmetric double well potential depending on the structure of the R and R' groups in the specific nitroxide.^{5,8,12,13} Typically, the barrier between the two wells is small and does not exceed a few kilocalories per mole, and at ambient temperature both wells are occupied with the NO bond continuously flipping between two minima. This conventional description of the PES associated with out of plane motion of the NO bond in DTBNO and 2,2,5,5-tetramethylpyrrolidine-N-oxyl has been recently challenged by Houriez and co-workers.¹³ On the basis of B3LYP/6-311+G(d,p) calculations, they proposed that the PES associated with the improper dihedral angle has a harmonic potential shape with a single minimum at $\theta = 0$ and exhibits a flat energy profile around the minimum. These findings are though quite surprising and contradict the geometry optimization results of Pavone et al.,⁸ Rinkevicius et al.,^{4,19} and Sinnecker et al.,⁷ which all show that DTBNO has a nonplanar ONRR' moiety at the equilibrium geometry. In order to settle this issue, we carried out a relaxed PES scan along the improper dihedral angle θ in DTBNO (see Figure 1) using the restricted open-shell MP2 method.⁵⁶ As one can see from Figure 1, our MP2 results clearly indicate that the PES associated with the motion of the NO bond in DTBNO has two wells and that the barrier between the wells is around 1.4 kcal/mol. On the basis of these findings, we will use a double well potential model for the PES associated with the NO bond out of plane motion in the interpretation of the nitrogen isotropic HFCC in DTBNO.

4.2. Structure and Local Solvent Dynamics of Di-*tert*-butyl Nitroxide in Aqueous Solution. According to the discussion in the previous section, the nitrogen isotropic HFCCs critically depend on the geometric structure of the ONRR' moiety, and an understanding of changes of these geometrical parameters upon solvation is critical for interpretation of the solvent induced shifts of the HFCCs. In previous works,^{8,13} both classical MD and Car–Parrinello MD have been employed to study the structure of DTBNO in aqueous solution. From these works, Car–Parrinello MD simulations by Pavone et al.⁸ are the most rigorous ones, as they avoid empirical parametrization of the classical force field of DTBNO used in classical MD¹³ and can thus model the internal molecular geometry of DTBNO and its solvent dependence more accurately compared to classical force-field-based approaches. However, the CP-MD method, being a pure density functional theory approach, neglects the dispersion interaction between the solute and solvent molecules as well as between the solvent molecules itself. In order to avoid this problem, we employed a hybrid CP-MD/MM method^{34,35} for simulation of the DTBNO in aqueous solution, which supposedly yields better results in describing the solute–solvent supermolecular structures, as it includes all important solute/solvent interactions—short-range repulsion, long-range dispersion, and electrostatic interactions. In the following, we will analyze the intramolecular geometry of DTBNO in solution and its solvation shell structure obtained in our CP-MD/MM simulation and compared with previous CP-MD results by Pavone et al.⁸ In Table 3, we list a few of the important internal geometrical parameters for DTBNO in aqueous solution obtained from our CP-MD/MM and previous CP-MD simulations. Interestingly, the internal geometry of DTBNO appears to be very similar in both CP-MD/MM and previous CP-MD simulations. However, the hybrid CP-MD/MM simulations predict a small increase in NO, NR, and NR' bond lengths (when compared to CP-MD results), and this naturally should be attributed to the increased solute–solvent interaction strength within the QM/MM framework. A similar solvent-induced

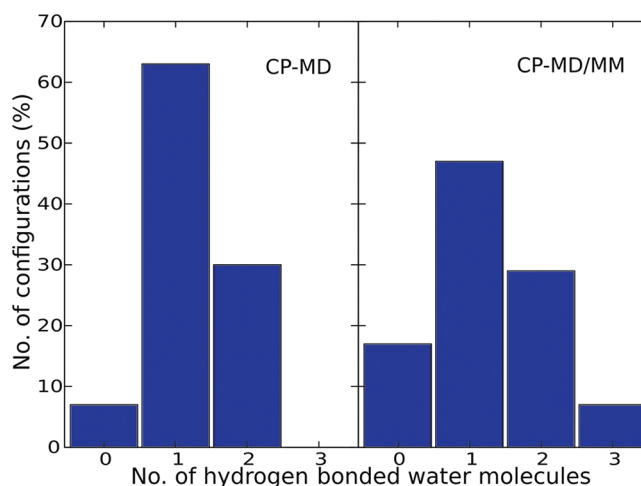
Table 3. Geometrical Structure of Di-*tert*-butyl Nitroxide in Aqueous Solution from Hybrid CP-MD/MM and CP-MD Simulations at Ambient Temperature

structural parameter ^a	CP-MD ^b	CP-MD/MM ^c	hydrogen bonds	CP-MD ^b	CP-MD/MM ^c
R_{NO} , ^d Å	1.31	1.32	number of H bonds	1.2	1.3
\bar{R}_{NR} , ^e Å	1.54	1.57	$R_{\text{NO}\cdots\text{O}}$, ^h Å	2.9	2.9
$\angle \text{RNR}'$, deg	127	130	$R_{\text{NO}\cdots\text{H}}$, ⁱ Å	1.9	1.9
$\angle \text{ONR}$, ^f deg	115	116	$\angle \text{O}\cdots\text{OH}$, ^j deg	15	18
$ \theta $, ^g deg	14.33	14.33			

^a Structural parameter values are averaged over CP-MD or CP-MD/MM trajectory. ^b Results are taken from work by Pavone et al.⁸ ^c This work. See Computational Details section. ^d R_{NO} is the NO bond length in DTBNO. ^e \bar{R}_{NR} is the average of NR and NR' bond lengths in DTBNO. ^f $\angle \text{ONR}$ is the average of angles ONR and ONR' in DTBNO. ^g $|\theta|$ is the absolute value of the improper dihedral angle, which describes NO bond motion out of the RNR' plane (see Figure 1). ^h $R_{\text{NO}\cdots\text{O}}$ is the averaged distance between the oxygen atom of the NO group in DTBNO and the oxygen atom of the hydrogen bonded water molecule. ⁱ $R_{\text{NO}\cdots\text{H}}$ is the averaged hydrogen bond length between the NO group in DTBNO and the water molecule. ^j $\angle \text{O}\cdots\text{OH}$ is the characteristic hydrogen bonding angle between oxygen of NO group in DTBNO and OH bond in hydrogen bonded water molecule.

stretching of bond lengths due to intermolecular hydrogen bonding has been reported in the literature previously in the case of acetone and phenol blue.^{57,58} It is interesting to note that the absolute value of the improper dihedral angle θ is the same in the case of CP-MD and CP-MD/MM simulations. The solute–solvent interactions as per the CP-MD and hybrid CP-MD/MM approaches are different with respect to dispersion and repulsion interaction energies; however, this brings only small changes in the internal molecular geometry around the ONRR' moiety. The NO bond length differs by 0.01 Å, while the average of NR and NR' bond lengths differs by 0.03 Å. Thus, overall, the geometrical structure of DTBNO obtained in our CP-MD/MM simulations is very similar to the one featured in CP-MD simulations of Pavone et al.⁸

Apart from the above-described structural parameters of DTBNO in aqueous solution, the local solvent structure around DTBNO also plays an important role in defining the solvent shift of nitrogen isotropic HFCC in this radical. Taking this into account, we focus on the structure of the first solvation shell since the solvent molecules in this shell are known to be most influential on the EPR spin Hamiltonian parameters of nitroxides, as they are in direct contact with the nitroxides through intermolecular hydrogen bonding and charge transfer. In the previous work by Pavone et al.,⁸ the first solvation shell was characterized in terms of the number of solvent molecules in hydrogen bonding with the DTBNO and the geometry of this intermolecular hydrogen bonding. In order to facilitate a direct comparison with their results, we have adopted the same geometrical parameters to define the hydrogen bonding.⁸ A solvent molecule is said to be hydrogen bonding with a solute molecule when the following three conditions are fulfilled: (a) $R_{\text{NO}\cdots\text{O}} \leq 3.5$ Å, (b) $R_{\text{NO}\cdots\text{H}} \leq 2.6$ Å, and (c) $\angle \text{O}\cdots\text{OH} \leq 30^\circ$. Figure 3 shows the populations of solute–solvent supermolecular structures in our CP-MD/MM simulation, where the number of intermolecular hydrogen bonding varies between 0 and 3. Interestingly, the average bond lengths between DTBNO and hydrogen-bonded water molecules ($R_{\text{NO}\cdots\text{O}}$ and $R_{\text{NO}\cdots\text{H}}$ in Table 3) appear to be the same for both, CP-MD and CP-MD/MM, approaches. Even the averaged characteristic angle, $\angle \text{O}\cdots\text{OH}$, is the same in both cases, suggesting that the average solute/solvent local structure is the same. However, as it is seen from Figure 3, the dynamical pictures of solute–solvent structure captured by these two approaches are quite different. The CP-MD approach could not find any solute–solvent structure with three hydrogen bonds between the solute and solvent molecules, while the hybrid CP-MD/MM

**Figure 3.** Hydrogen bonding patterns of di-*tert*-butyl nitroxide in aqueous solution at ambient temperature obtained from CP-MD (Pavone et al.⁸) and CP-MD/MM (this work) simulations.

approach predicts the population of such structures to be around 7%. Moreover, the solute–solvent structures with zero hydrogen bonding are more populated in the case of hybrid CP-MD/MM simulations. Both approaches predict an almost equal population for the solute–solvent structures with two hydrogen bonds. The average number of hydrogen bonds in the solute–solvent structures is 1.2 in the case of the CP-MD approach, while it is 1.3 in the case of the hybrid CP-MD/MM approach, and so the latter approach predicts a slightly larger coordination number for hydrogen-bonded solvent molecules. Overall, the results suggest that the average (intermolecular) geometry of solute–solvent structure is the same as obtained from both approaches, while the dynamic description of this is very different in both cases. Here, we notice that a similar situation has been encountered in modeling of the copper dication solvation in water,⁵⁹ where the CP-MD/MM approach predicted a six coordinated structure while the CP-MD-approach-based simulations predicted a five coordinated structure, and where the structure from the latter approach has been discussed to be an artifact due to the neglect of dispersion interactions. In the forthcoming section, we will discuss the nitrogen isotropic HFCC obtained by averaging over a set of snapshots from the above-described CP-MD/MM simulations and investigate the influence of the local solvent

environment around the solute on this EPR spin Hamiltonian parameter.

4.3. Isotropic Hyperfine Coupling Constant of Nitrogen in Di-*tert*-butyl Nitroxide in Aqueous Solution. After describing the isotropic HFCC of nitrogen in DTBNO from the perspective of the DFT-RU approach and analyzing the geometrical structure of DTBNO in aqueous solution, let us turn to the main topic of this work, namely, the capability of the DFT-RU/MM approach to predict the nitrogen isotropic HFCC in DTBNO solvated in water. First, we will discuss the dependence of the nitrogen isotropic HFCC on the description of the MM region and, later on, compare our results with experimental data^{51–53} as well as findings from previous works^{8,13} which exploited the integrated approach.²⁰

In Table 4, we give the isotropic HFCC of nitrogen in DTBNO in water computed by DFT-RU/MM. First, we consider the simplest possible model of the QM region, which contains only the DTBNO radical and with all water molecules within a 20 Å radius moved to the MM region. As expected, according to our analysis of the nitrogen isotropic HFCC in DTBNO, the direct spin density contribution constitutes only 14.5–17.2% of the total a_N value, averaged over 86 snapshots, depending on the MM region treatment and on the exchange–correlation functional (see Table 4), and the remaining contribution to the a_N value originates from the spin polarization contribution. We point out that such a a_N composition, with some minor variations in terms of direct spin density and spin polarization contributions, is observed from most of the snapshots used in our DFT-RU/MM calculations, and only for a few snapshots in which the improper dihedral angle θ is large does the direct spin density contribution to a_N turn out to be larger than the spin polarization contribution.

After establishing the composition of the nitrogen isotropic HFCC in DTBNO, we turn to its solvent shift. As one can see from Table 4, the DFT-RU/MM approach predicts the solvent shift of a_N , depending on the force field used in the water MM region, to be 1.98–2.65 G at the B3LYP level and 2.19–2.94 G at the PBE0 level, respectively. For the crudest description by the MM-0 force field, which contains only point charges, the calculations at the B3LYP level predict the solvent shift to be 1.98 G or around 15% of the a_N value in a vacuum. A slightly larger solvent shift, equal to 2.19 G, is obtained at the PBE0 level. Here, we would like to point out that the vibrational contribution from internal DTBNO dynamics to a_N in our DFT-RU calculations is estimated to be 0.21 at the B3LYP level and, in agreement with the results of Pavone et al.,⁸ is much smaller than the solvent shift. Exchanging MM-0 with the more elaborate MM-1 force field, which in addition to point charges also contains the isotropic water polarizability (see Table 1), turns out to have little effect on nitrogen isotropic HFCC. This is due to the fact that the MM-0 force field point charge values are constructed to mimic the solvent environment and account implicitly for polarization by artificially increasing the values of the point charges used to describe the water molecules. Further improvement of the force field from MM-1 to MM-2, which introduces distributed anisotropic polarizability tensors in the description of water molecules in the MM region, leads to a significant increase in the solvent solvent shift of a_N by almost 20–23% depending on the exchange–correlation functional compared to the one obtained in the calculations with the MM-0 or MM-1 force fields. Finally, the last two force fields, MM-3 and MM-4, which go beyond the point charge approximation and also provide more refined water polarizability by using more expansion points, give almost identical HFCC values (see Table 4) and predict the a_N

Table 4. Isotropic Hyperfine Coupling Constant of Nitrogen of the Di-*tert*-butyl Nitroxide in Water Computed with DFT/MM Methods Employing Two QM Regions: One Which Includes Only Di-*tert*-butyl Nitroxide and Another Which Includes Di-*tert*-butyl Nitroxide with Two Water Molecules Hydrogen Bonded to the NO Bond^{a,b}

method	DTBNO			DTBNO+2H ₂ O		
	$a_N^{\text{den}}, \text{G}$	$a_N^{\text{pol}}, \text{G}$	a_N, G	$a_N^{\text{den}}, \text{G}$	$a_N^{\text{pol}}, \text{G}$	a_N, G
B3LYP	2.26	10.86	13.12	2.28	12.14	14.43
B3LYP/MM-0	2.34	12.75	15.10	2.32	12.90	15.23
B3LYP/MM-1	2.33	12.82	15.15	2.32	12.90	15.23
B3LYP/MM-2	2.34	13.21	15.55	2.32	13.08	15.40
B3LYP/MM-3	2.36	13.38	15.74	2.33	13.14	15.47
B3LYP/MM-4	2.36	13.41	15.77	2.32	13.15	15.48
PBE0	2.24	11.32	13.56	2.29	12.75	15.04
PBE0/MM-0	2.36	13.39	15.75	2.34	13.59	15.93
PBE0/MM-1	2.35	13.46	15.81	2.34	13.59	15.93
PBE0/MM-2	2.36	13.91	16.27	2.35	13.77	16.12
PBE0/MM-3	2.39	14.08	16.47	2.35	13.84	16.19
PBE0/MM-4	2.39	14.11	16.50	2.35	13.86	16.21
PBE0 ^c			13.1			
PBE0/PCM ^c			12.6			14.7
PBE0 ^d			14.2			
PBE0/ESPF ^d			16.7			
exptl. ^e			16.75 ± 0.04			16.75 ± 0.04
exptl. ^f			17.18 ± 0.01			17.18 ± 0.01
exptl. ^g			17.16 ± 0.01			17.16 ± 0.01

^a Nitrogen isotropic HFCC, a_N , has been computed as the average over 86 snapshots extracted from the CP-MD/MM trajectory. ^b All DFT-RU calculations have been performed using the hybrid B3LYP exchange–correlation functional and Huz-IIIu3 basis set. ^c Unrestricted DFT results obtained for a vacuum and DFT/PCM results for water solution by Pavone et al.⁸ Calculations were carried out using the EPR-II basis set, and results are averaged over snapshots extracted from the CP-MD trajectory. ^d Unrestricted DFT results obtained for a vacuum and DFT/ESPF results for water solution by Houriez et al.¹³ Calculations were carried out using 6-31G+(d,p) basis set, and results are averaged over snapshots extracted from the classical MD trajectory. ^e Experimental data taken from work by Kawamura et al.⁵¹ EPR spectra measurements have been carried out at ambient temperature. ^f Experimental data taken from work by Knauer and Napie.⁵³ EPR spectra measurements have been carried out at ambient temperature. ^g Experimental data taken from work by Griffith et al.⁵² EPR spectra measurements have been carried out at ambient temperature.

solvent shift at the B3LYP level to be 2.62 or 2.65 G, which is around 8% larger than the solvent shift predicted using the MM-2 force field. Similar results are also obtained with the PBE0 functional, which gives a solvent shift for MM-3 and MM-4 force fields of 2.91 and 2.94 G, respectively. Taking these results into account, we conclude that the solvent shift of the nitrogen isotropic HFCCs in DTBNO is highly dependent on the quality of the force field used for the description of the MM region; independently of the exchange–correlation functional used, its value increases going from the MM-0 to the MM-4 force field in the order MM-0 ≈ MM-1 < MM-2 < MM-3 ≈ MM-4. Thus, the MM-3 force field provides a converged description of the water potential in the DFT-RU/MM calculations in which the QM region is limited to the solute itself, and force fields designed

using the same procedure can be recommended for general computations of hyperfine coupling constants in more complex environments.

After discussing the quality of the MM description in the case of the simplest QM model, which contains only one DTBNO molecule, we attend to the results obtained with a more complex QM region model, which includes the DTBNO radical and the two closest water molecules to the NO bond of DTBNO. This model of the QM region is frequently employed in DFT/PCM studies of various properties of nitroxides,⁵ as it explicitly accounts for the hydrogen bonding between the nitroxide and solvent molecules and in turn allows one to study nitroxides in protic solvents within the DFT/PCM method. Thus, by comparing the DFT-RU results obtained using the two QM region models, we estimate that the hydrogen-bonding-induced solvent shift of a_N in DTBNO is equal to 1.31 G at the B3LYP level and to 1.48 G at the PBE0 level, respectively. The remaining part of the solvent shift originates from the water molecules in the bulk, and according to our DFT-RU/MM results (see Table 4), this is, depending on the force field used, between 0.8 and 1.05 G for the B3LYP functional and between 0.89 and 1.17 G for the PBE0 functional. A closer inspection of the DFT-RU/MM results reveals that the solvent shift associated with water molecules in the MM region shows a similar behavior with respect to force field quality as in the above-discussed case and increases in the following order independently of the exchange–correlation functional used in calculations: MM-0 \approx MM-1 < MM-2 < MM-3 \approx MM-4. However, differences between results obtained with different force fields are less pronounced, since the water molecules which are most sensitive to the quality of the force field in this case are included in the QM region.

The QM region models, namely the one which includes only DTBNO and the other which includes “DTBNO+2 waters”, behave rather similarly in DFT-RU/MM calculations, and differences between two models for the HFCCs do not exceed 0.29 G for both exchange-correlation functionals used in this work. For the lower quality force fields, MM-0 and MM-1, the second QM region model, which includes water molecules explicitly, gives rise to slightly larger a_N values compared to the first QM region model, while the opposite holds for the other force fields used in this work. Here, we would like to point out that the differences between the two QM region models from the perspective of the overall accuracy of the nitrogen isotropic HFCC, obtained using the DFT-RU/MM approach, are rather small, and usage of the simpler QM region model introduces an error similar in size to the one obtained truncating the basis set from Huz-IVsu4 to Huz-IIIsu3 (see Table 2). Comparing the DFT-RU/MM results obtained for both QM region models (see Table 4), it is evident that the changes in the description of the waters hydrogen-bonded to DTBNO from quantum to classical affects only the changes of the spin polarization contribution to a_N , while the direct spin density contribution remains almost unchanged. It becomes clear that the question of which combination of the QM region model and force field is sufficiently accurate for routine calculations of nitrogen isotropic HFCCs in nitroxides and spin labels is nontrivial. From the conceptual point of view, the QM region model, which includes DTBNO and two hydrogen-bonded water molecules, and a force field parametrized in terms of distributed multipoles and anisotropic polarizability tensors, is the most suitable combination for DFT-RU/MM calculations. As we can see from Table 4, a converged description of the nitrogen isotropic HFCC in DTBNO for this QM region model

is obtained already for the MM-3 force field. However, explicit inclusion of solvent molecules into the QM region is highly undesirable, as it significantly increases the computational cost and limits the applicability of the DFT-RU/MM method to complex environments, like cellular membranes or ion channels. Therefore, for practical applications of the DFT-RU/MM method, it would be advantageous to limit the QM region to the solute itself. A closer inspection of Table 4 reveals that combining the simpler QM region model, which includes only DTBNO, with the MM-2 force field gives a_N values close to the converged ones computed using the more complex QM region model and MM-3 force field. On the basis of these findings, we can recommend limiting the QM region in practical DFT-RU/MM calculations to the nitroxide of interest and treating the environment with a force field of similar quality to that of the MM-2 force field for water. This recommendation relies on error cancellation between the description of the bulk part of the solvent environment and the hydrogen-bonded solvent molecules, and its validity must be tested further and especially for nonaqueous solutions.

After settling the issues with suitable QM region models and force fields for DFT-RU/MM calculations, let us turn our attention to the solvent-induced shift of this quantity. For both QM region models, our DFT-RU/MM calculations, independently of the force field used, show that the direct spin density contribution to a_N is almost not at all affected by the solvation of DTBNO and that the solvent shift of the isotropic HFCC arises solely from the spin polarization contribution. This statement holds for most of the snapshots extracted from the CP-MD/MM trajectory, but for some snapshots in which the improper dihedral angle θ (see Figure 1) is large, the direct spin density contribution to a_N can also exhibit a solvent shift. However, instantaneous configurations of DTBNO with large improper dihedral angles θ are rarely encountered in CP-MD/MM simulations at ambient temperature due to the steep PES for this degree of freedom (see Figure 1), and therefore they will have little impact on the overall behavior and size of the solvent shift of a_N .

After establishing the origin of the solvent shift of the nitrogen isotropic HFCC in DTBNO solvated in water, let us turn to the dependence of this shift on the solvent environment. From results presented in Table 4, we can estimate that the solvent shift of a_N is around 2.35 G at the B3LYP level (QM region model: “DTBNO+2 waters” and MM-3 force field), where around 56% of the solvent shift originates from direct hydrogen bonding and the remaining part from bulk solvent. At the PBE0 level, the solvent shift is slightly larger and is around 2.63 G, but the relative size of hydrogen bonding and bulk solvent contributions remains similar to the B3LYP case. Unfortunately, a more detailed explanation of the spin polarization contribution to a_N upon solvation is not easily achieved, as it turns out that many triplet orbital rotations contribute to the electron density relaxation. Finally, concluding the discussion of solvent shift of the nitrogen isotropic HFCC in DTBNO, we point out that our calculation results conform with the conventional explanation of this solvent shift as a consequence of the preferential stabilization of the zwitterionic form of the ONRR' moiety. However, we would like to emphasize that the origin of the larger nitrogen isotropic HFCCs in the zwitterionic form is due to a more effective spin polarization and not because of an increase in the direct spin density contribution as has been suggested previously.

To complete our discussion of nitrogen isotropic HFCCs in DTBNO solvated in water, we compare the results obtained using our DFT-RU/MM approach with experimental data and

results from previous studies with the integrated approach. As we can see from Table 4, independently of the QM region model and force field used, the DFT-RU/MM approach underestimates the experimentally measured a_N by 1.4–2.1 G at the B3LYP level and by 0.6–1.4 G at the PBE0 level. Thus, the PBE0 functional predicts nitrogen HFCCs in radicals with better accuracy than the commonly used B3LYP functional, as was already noted in previous works devoted to HFCCs.^{17,48} Taking into account the strong dependence of the nitrogen isotropic HFCCs on the exchange–correlation functional and basis set used in specific DFT calculations, our DFT-RU/MM results compare quite well with data from the unrestricted DFT/PCM calculations of Pavone et al.⁸ and unrestricted DFT/MM calculations of Houriez et al.¹³ On the basis of this set of results, we can confirm that the DFT-RU/MM approach provides more accurate results than conventional unrestricted DFT-based approaches and that it thus can be recommended for studies of nitroxides in various complex environments.

5. CONCLUSIONS

In the present work, we have taken a critical step toward realistic modeling and design of molecular spin probes in solution or confined environments. Our work combines up-to-date electronic structure theory for open shell systems with new developments of multiscale modeling techniques for solvent interactions. The former encompasses a density functional theory that “solves” the spin problem in that the reference state can be maintained spin-uncontaminated, thus allowing one to obtain a strict one-to-one mapping between electron density and spin state of the molecule, yet proper spin polarization is introduced in the calculation of the property. This, the so-called restricted–unrestricted DFT approach earlier introduced by the authors, has shown good performance in prediction of spin Hamiltonian parameters, including the hyperfine coupling constant addressed in the present work. We then used a QM/MM response theory with a full account of intermolecular interaction for the quantum mechanical property of the solute.

A demonstration is given for the nitrogen isotropic hyperfine coupling constant in *di-tert*-butyl nitroxide solvated in water for which we in detail investigated the representation of the QM and MM regions required for an accurate prediction of this constant, thus the need for additional hydrogen-bonded molecules treated quantum mechanically and the degree of granularity in the charge and polarizability distributions of the MM water molecules. It is found that an accurate MM force field reduces the need to include explicit water molecules in the QM region, something that significantly decreases computation cost and widens the applicability to more complex environments. The DFT-RU/MM technique allowed us to pinpoint the origin of the solvent shift in great detail: Part of the solvent dependence originates in that the nitrogen isotropic HFCC critically depends on the geometrical structure of the ONRR' moiety and that these geometrical parameters change rather significantly upon solvation. Moreover, only the direct spin density contribution to the nitrogen isotropic HFCC exhibits a similar dependence on NO bond out of plane motion to that of the isotropic HCCC itself, while the spin polarization contribution behaves very differently and slowly decreases with the increase of the absolute value of the improper dihedral angle between the NO bond and NRR' moiety. The prominent role of spin polarization is an important finding, which gives the interpretation of the shift of nitrogen isotropic HFCC

upon solvation as an adjustment to the change of spin density relaxation, i.e., spin polarization change caused by solute interaction with solvent molecules.

We find that the RU-DFT/MM methods allow one to calculate nitroxide HFFCs with good accuracy at the same time as it provides interpretation to the dependence on electron and geometrical structure and solvent environment. In particular, the ability of the method to separate the spin polarization from the direct spin density contributions allows one to disentangle the structure–property relations for these constants. This in turn is an important aspect when turning to more complex, confined environments, like protein cavities or membrane channels, where application-specific design of the nitroxide spin labels with optimal performance is desirable.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rinkevic@theochem.kth.se.

ACKNOWLEDGMENT

This work has been partially funded by the EU Commission (contract INFISO-RI-261523) under the ScalaLife collaboration and has been supported by a grant from the Swedish Infrastructure Committee (SNIC) for the project “Multiphysics Modeling of Molecular Materials”, SNIC 022/09-25. J.K. thanks The Danish Councils for Independent Research and the Lundbeck Foundation for financial support.

REFERENCES

- (1) Owenius, R.; Engström, M.; Lindgren, M.; Huber, M. *J. Phys. Chem. A* **2001**, *105*, 10967–10977.
- (2) Engström, M.; Vaara, J.; Schimmelpennig, B.; Ågren, H. *J. Phys. Chem. B* **2002**, *106*, 12354–12360.
- (3) D'Amore, M.; Improta, R.; Barone, V. *J. Phys. Chem. A* **2003**, *107*, 6264–6269.
- (4) Rinkevicius, Z.; Telyatnyk, L.; Vahtras, O.; Ruud, K. *J. Chem. Phys.* **2004**, *121*, S051–S060.
- (5) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231–1254.
- (6) Neugebauer, J.; Louwse, M. J.; Belanzoni, P.; Wesolowski, T. A.; Baerends, E. J. *J. Chem. Phys.* **2005**, *123*, 114101.
- (7) Sinnecker, S.; Rajendran, A.; Klamt, A.; Diedenhofen, M.; Neese, F. *J. Phys. Chem. A* **2006**, *110*, 2235–2245.
- (8) Pavone, M.; Cimino, P.; De Angelis, F.; Barone, V. *J. Am. Chem. Soc.* **2006**, *128*, 4338–4347.
- (9) Barone, V.; Brustolon, M.; Cimino, P.; Polimeno, A.; Zerbetto, M.; Zoleo, A. *J. Am. Chem. Soc.* **2006**, *128*, 15865–15873.
- (10) Carlotto, S.; Cimino, P.; Zerbetto, M.; Franco, L.; Corvaja, C.; Crisma, M.; Formaggio, F.; Toniolo, C.; Polimeno, A.; Barone, V. *J. Am. Chem. Soc.* **2007**, *129*, 11248–11258.
- (11) Pavone, M.; Cimino, P.; Crescenzi, O.; Sillanpää, A.; Barone, V. *J. Phys. Chem. B* **2007**, *111*, 8928–8939.
- (12) Houriez, C.; Ferré, N.; Masella, M.; Siri, D. *J. Chem. Phys.* **2008**, *128*, 244504.
- (13) Houriez, C.; Ferré, N.; Siri, D.; Masella, M. *J. Phys. Chem. B* **2009**, *113*, 15047–15056.
- (14) Houriez, C.; Ferré, N.; Masella, M.; Siri, D. *THEOCHEM* **2009**, *898*, 49–55.
- (15) Pavone, M.; Biczysko, M.; Rega, N.; Barone, V. *J. Phys. Chem. B* **2010**, *114*, 11509–11514.
- (16) Barone, V.; Cimino, P.; Pedone, A. *Magn. Reson. Chem.* **2010**, *48*, S11–S22.
- (17) Herмосilla, L.; García de la Vega, J. M.; Sieiro, C.; Calle, P. *J. Chem. Theory Comput.* **2011**, *7*, 169–179.

- (18) Ikryannikova, L. N.; Ustyniuk, L. Y.; Tikhonov, A. N. *Magn. Reson. Chem.* **2010**, *48*, 337–349.
- (19) Rinkevicius, Z.; Murugan, N. A.; Kongsted, J.; Aidas, K.; Steindal, A. H.; Ågren, H. *J. Phys. Chem. B* **2011**, *115*, 4350–4358.
- (20) Crescenzi, O.; Pavone, M.; De Angelis, F.; Barone, V. *J. Phys. Chem. B* **2005**, *109*, 445–453.
- (21) Asher, J. R.; Kaupp, M. *Theor. Chem. Acc.* **2008**, *119*, 477–487.
- (22) Pauwels, E.; Declerck, R.; Verstraelen, T.; De Sterck, B.; Kay, C. W. M.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. B* **2010**, *114*, 16655–16665.
- (23) Sezer, D.; Freed, J. H.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 5755–5767.
- (24) Stendardo, E.; Pedone, A.; Cimino, P.; Menziani, M. C.; Crescenzi, O.; Barone, V. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11697–11709.
- (25) Olsen, J. M.; Aidas, K.; Kongsted, J. *J. Chem. Theory Comput.* **2010**, *6*, 3721–3734.
- (26) Rinkevicius, Z.; Telyatnyk, L.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2004**, *121*, 7614–7623.
- (27) Oprea, C. I.; Telyatnyk, L.; Rinkevicius, Z.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2006**, *124*, 174103.
- (28) Rinkevicius, Z.; de Almeida, K. J.; Vahtras, O. *J. Chem. Phys.* **2008**, *129*, 064109.
- (29) Fernández, B.; Jørgensen, P.; Byberg, J.; Olsen, J.; Helgaker, T.; Jensen, H. J. A. *J. Chem. Phys.* **1992**, *97*, 3412–3419.
- (30) Fernández, B.; Christiansen, O.; Bludsky, O.; Jørgensen, P.; Mikkelsen, K. V. *J. Chem. Phys.* **1996**, *104*, 629–635.
- (31) DALTON program. www.daltonprogram.org (accessed Aug 2011).
- (32) Møgelhøj, A.; Aidas, K.; Mikkelsen, K. V.; Sauer, S. P. A.; Kongsted, J. *J. Chem. Phys.* **2009**, *130*, 134508.
- (33) Aidas, K.; Mikkelsen, K. V.; Mennucci, B.; Kongsted, J. *Int. J. Quantum Chem.* **2011**, *111*, 1511–1520.
- (34) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.
- (35) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.
- (36) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (37) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (38) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (39) Jørgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (40) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.
- (41) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (42) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (43) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (44) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (45) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (46) Hermosilla, L.; Calle, P.; García de la Vega, J. M.; Sieiro, C. *J. Phys. Chem. A* **2005**, *109*, 1114–1124.
- (47) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751–764.
- (48) Barone, V.; Cimino, P. *Chem. Phys. Lett.* **2008**, *454*, 139–143.
- (49) van Wüllen, C. *Die Berechnung magnetischer Eigenschaften unter Berücksichtigung der Elektronkorrelation: Die Multikonfigurations-Verallgemeinerung der IGLO-Methode*. Ph.D. thesis, Ruhr-Universität, Bochum, Germany, 1992.
- (50) Lantto, P.; Vaara, J.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 5998–6009.
- (51) Kawamura, T.; Matsunami, T.; Yonezawa, T. *Bull. Chem. Soc. Jpn.* **1967**, *40*, 1111–1115.
- (52) Griffith, O. H.; Dehlinger, P. J.; Van, S. P. *J. Membr. Biol.* **1974**, *15*, 159–192.
- (53) Knauer, B. R.; Napier, J. J. *J. Am. Chem. Soc.* **1976**, *98*, 4395–4400.
- (54) Stone, A. J. *Mol. Phys.* **1963**, *6*, 509–515.
- (55) Stone, A. J. *Mol. Phys.* **1964**, *7*, 311–316.
- (56) Aikens, C. M.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. *J. Chem. Phys.* **2006**, *124*, 014107.
- (57) Röhrig, U. F.; Frank, I.; Hutter, J.; Laio, A.; VandeVondele, J.; Rothlisberger, U. *ChemPhysChem* **2003**, *4*, 1177–1182.
- (58) Murugan, N. A.; Rinkevicius, Z.; Ågren, H. *J. Phys. Chem. A* **2009**, *113*, 4833–4839.
- (59) de Almeida, K. J.; Murugan, N. A.; Rinkevicius, Z.; Hugosson, H. W.; Vahtras, O.; Ågren, H.; Cesar, A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 508–519.
- (60) Ahlström, P.; Wallqvist, A.; Engström, S.; Jönsson, B. *Mol. Phys.* **1989**, *68*, 563–581.
- (61) Helgaker, T.; Jaszunski, M.; Ruud, K.; Górska, A. *Theor. Chem. Acc.* **1998**, *99*, 175–182.
- (62) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (63) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

Double-Hybrid Density Functionals Provide a Balanced Description of Excited 1L_a and 1L_b States in Polycyclic Aromatic Hydrocarbons

Lars Goerigk and Stefan Grimme*

Organisch-Chemisches Institut der Universität Münster, Corrensstraße 40 D-48149, Münster, Germany

 Supporting Information

ABSTRACT: The time-dependent density functional theory (TD-DFT) double-hybrid methods TD-B2-PLYP and TD-B2GP-PLYP are applied to five linear and 12 nonlinear polycyclic aromatic hydrocarbons. The absolute errors compared to experiment for the two lowest-lying 1L_a and 1L_b excited states are evaluated and it is also tested whether the energetic order of those states and their energy difference is reproduced correctly. The results are compared to published CC2, global hybrid, and long-range corrected hybrid TD-DFT results. The two double-hybrids outmatch the other methods in terms of absolute and relative accuracy without an empirical adjustment of parameters. Although of different electronic character, both types of states are described on an equal footing by the double-hybrids. Particularly, the B2GP-PLYP functional yields very good results, which is in accordance with previous benchmarks.

1. INTRODUCTION

The theoretical description of the two lowest lying $\pi \rightarrow \pi^*$ (1L_a and 1L_b)¹ transitions in polycyclic aromatic hydrocarbons (PAHs) has sporadically gained particular interest in the past eight years. They are considered as being prototypical for low-lying excited states in practically important organic dyes. In 2003, Grimme and Parac showed^{2,3} that conventional time-dependent density functional theory (TD-DFT)^{4–7} methods severely fail in correctly describing the 1L_a states in linear² and nonlinear³ PAHs. Not only were the absolute errors larger than usually expected for valence excitation energies obtained with TD-DFT (1L_a excitations were highly underestimated), but also the molecule size-dependence of the errors was larger than found for wave function based approaches. Moreover, the energy difference between the 1L_a and 1L_b states and their relative order was very often wrong. On the basis of a semiempirical TD-PPP treatment,³ these errors were related to a theoretically estimated ionicity measure of the 1L_a states, which is higher for more delocalized π -systems. It is expected that a certain amount of ionicity may generally play a role in extended organic chromophores. As a first remedy of this problem, one could increase the amount of Fock exchange in global hybrid functionals, which generally leads to a blue-shift and smaller errors for 1L_a type states; however, more covalent (nonionic) states are also affected, and consequently, their energies are overestimated. Several studies showed that empirically about 40% nonlocal Fock exchange for larger and about 25% for smaller chromophores represents a good compromise.^{8–10} Another possibility to increase the results for extended conjugated π -chromophores is the DFT/MRCI approach,¹¹ as shown by Marian and Gilka (among others also for the series of linear acenes).¹² However, for large systems with many states, applying DFT/MRCI becomes quickly unfeasible, and one wishes a robust single-reference TD-DFT method to do the same good job.

Range-separated or long-range corrected (LRC) functionals were also thoroughly assessed for valence excitations and showed

promising results.^{9,13,14} In 2010, Wong and Hsieh evaluated LRC functionals for the same series of linear oligoacenes as Grimme and Parac in 2003 and observed a substantial improvement for the 1L_a states compared to global hybrids.¹⁵ Very recently, Richard and Herbert also investigated nonlinear PAHs with LRC functionals and furthermore analyzed the nature of the 1L_a states by various theoretical methods.¹⁶ Those studies revealed that LRC functionals indeed showed smaller errors than global hybrids for the 1L_a states. However, 1L_b states were not computed more accurately; in fact, their description sometimes worsened. Moreover, the results depended strongly on the long-range correction parameter used for the range-separated exchange-operator. Its optimal value for 1L_a states in PAHs is not necessarily the best for other chromophore classes (or state types). Richard and Herbert suggested that 1L_a excitations have a partial charge-transfer (CT) character. However, standard diagnosis tools did not show conclusive evidence for it. Furthermore, the extent of ionicity was shown to be dependent on the method with which it was estimated. Thus, the nature of 1L_a states in PAHs is still not fully revealed.

As mentioned above, also excited states of other extended chromophores can have some kind of 1L_a or 1L_b character. From a practical and also theoretical point of view we regard it as rather tedious and questionable to adjust the range-separation factor of an LRC functional for a certain class of chromophores. Particularly for the application to hitherto unknown systems, this procedure would be difficult to apply without additional computational effort (for recent advances in determining system-dependent range-separation factors in a nonempirical way, see refs 17–19). It seems desirable to have a method that could describe states with varying $^1L_a/^1L_b$ character equally well.

Time-dependent double-hybrid density functionals (TD-DHDFs) proved to be promising candidates for a balanced

Received: June 7, 2011

Published: September 02, 2011

and accurate description of low-lying excited states.²⁰ It was shown that the TD-B2-PLYP^{20,21} and TD-B2GP-PLYP^{22,23} methods can outmatch LRC functionals and the (SCS-)CC2 methods for large dyes and that they are the most accurate TD-DFT methods in excited state benchmarks.^{8,9,23} In 2009, TD-B2-PLYP was successfully applied to theoretical electronic circular dichroism spectroscopy.²⁴ In one case, a chirally substituted phenyl chromophore was examined and TD-B2-PLYP could not only reproduce the excitation energies of the 1L_a and 1L_b states but also their energy difference very well.

Herein, we will further investigate how well double-hybrid density functionals (DHDFs) can describe the 1L_a and 1L_b transitions in PAHs. First, five acenes from naphthalene to hexacene will be examined, followed by 12 nonlinear PAHs. The TD-B2-PLYP and TD-B2GP-PLYP methods will be tested and compared with previously published TD-DFT, DFT/MRCI, and CC2 values. We will not only assess the error in the excitation energies compared to experiment but will also see how well the 1L_a - 1L_b splitting and the state order is described.

2. COMPUTATIONAL DETAILS AND THEORY

Double-hybrid density functionals (DHDFs) consist of two parts: a hybrid-GGA-like portion and a perturbative correction (see ref 20 for a thorough introduction). First, a standard linear response TD-DFT treatment is carried out for the hybrid part of the functional. In the following, we will denote these portions as B2-LYP or B2GP-LYP, respectively. B2-LYP has 53% Fock exchange, and the latter has 65%. On the basis of the resulting TD-DFT excitation energies and the ground state orbitals, a perturbative CIS(D)²⁵ type treatment is then carried out. The resulting correlation energies are scaled as for the ground state by 0.27 (B2-PLYP) and 0.36 (B2GP-PLYP), respectively, and combined with the hybrid result, yielding the final DHDF excitation energy. In the following discussions, both the resulting energies from the hybrid portions and the full DHDF energies will be considered to separately evaluate the influence of Fock exchange and perturbative correlation on the excitation energies.

For the present work, BH-LYP,²⁶ B2-PLYP, and B2GP-PLYP calculations were performed with a modified version of TURBOMOLE 5.9.^{27–30} The perturbative CIS(D)-type corrections for the double-hybrids were obtained with our group's own program, RICC,³¹ by making use of the resolution of the identity (RI) approach.³² For all calculations, Dunning's cc-pVTZ basis set was applied,³³ which yields results close to the basis set limit for the investigated property. Auxiliary basis functions for the RI approximation were taken from the TURBOMOLE library.³⁴ Geometries for all systems will be taken from the work by Richard and Herbert,¹⁶ and excitation energies generally refer to vertical excitations. Results for other methods were also taken from previous studies based on the cc-pVTZ basis set (except for DFT/MRCI results, which are based on TZVP calculations).^{2,12,16} The prefix "TD" in front of a functional's name will be skipped in the following. The SCF/TD steps were performed with the TURBOMOLE grid m4³⁴ and an SCF energy convergence criterion of $10^{-7} E_h$.

Before continuing with the discussion of the results, we want to make a comment on the application of DHDFs with standard program codes. Even though many programs allow double-hybrid calculations for electronic ground states, they do not necessarily also provide a combination with the CIS(D) type correction, when excited states are calculated. During recent

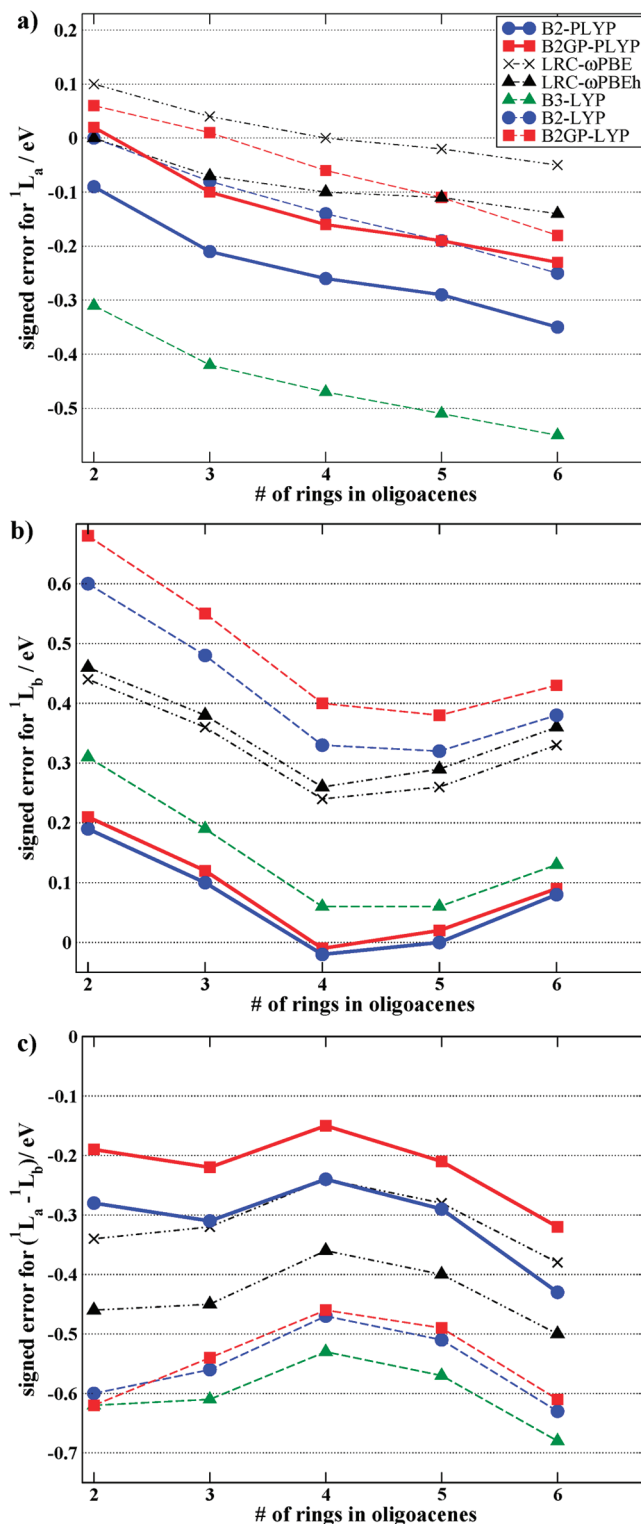


Figure 1. Signed errors (difference between theory and experiment) for various functionals for the 1L_a states (a), the 1L_b states (b), and the energy splittings between both states (c) of five polyacenes (from naphthalene to hexacene). All results are based on calculations with the cc-pVTZ basis. The results for B3-LYP and the two LRC methods were taken from ref 16. Theoretically corrected, experimental reference values were taken from ref 2.

activities as reviewers we became aware that allegedly the TD-B2PLYP method was applied, but in fact, the CIS(D) correction

seemed to have been excluded. However, usage of the perturbative CIS(D) correction as outlined in the first TD-B2-PLYP publication²⁰ is crucial. We encourage users and developers of standard quantum chemical software to take this fact into account in the future in order to prevent confusion and wrong usage.

3. RESULTS AND DISCUSSION

3.1. Linear Polyacenes. In the following discussion, results for the five acenes from naphthalene to hexacene will be considered, first. Theoretically corrected (i.e., back-corrected for vibrational effects) experimental reference values were taken from ref 2. Figure 1a shows the signed errors for various methods for the 1L_a states. The B3-LYP and LRC functionals' values were taken from ref 16. All curves show a similar behavior with system size, i.e., increasing errors for larger acenes. B3-LYP suffers from a severe underestimation from about -0.3 to -0.55 eV (see also Table S1 in the Supporting Information for the excitation energies of all tested methods). Increasing the amount of Fock exchange leads to a blue-shift, as can be seen for B2-LYP and B2GP-LYP. The errors for the latter range between 0.1 and -0.05 eV. Adding the perturbative correlation contribution leads to a red-shift of the results compared to the hybrid-GGA parts. B2-PLYP errors now range between -0.1 and -0.35 eV. B2GP-PLYP yields very good results within the accuracy of the reference values with an error range from about 0 to -0.2 eV. The LRC- ω PBEh method, as tested by Richard and Herbert, performs similarly to B2GP-PLYP; LRC- ω PBE has the smallest error range of the methods in Figure 1a, confirming previous conclusions on LRC functionals applied to 1L_a states.^{15,16} Note, that during the review process another work on range-separated functionals on PAHs was published, in which also LRC methods worked well for linear acenes.¹⁹

The picture given above changes when analyzing the errors for the 1L_b states (Figure 1b and Table S2 in the Supporting Information). All methods on average overestimate the excitation energies. Hybrid-functionals show larger errors with increasing amounts of nonlocal exchange. The long-range corrected methods are not much better than B2-LYP and worse than B3-LYP. The smallest errors are observed for the two DHDFs. This finding indicates that DHDFs more adequately describe correlation effects for the more multiconfigurational 1L_b states because the CIS(D) correction accounts for higher than single excitations which are formally absent in TDDFT and which are more important for 1L_b than for 1L_a states. Both DHDFs yield very similar results and have an error range between 0.2 and 0 eV.

Finally, the energy difference between the 1L_a and 1L_b states is considered and compared to the experimental values (Figure 1c). Even if both states had rather large absolute errors, their splitting could be still acceptable due to systematic errors. This finding would be particularly important for theoretical spectroscopy, for which it is sometimes sufficient to apply constant shifts to the spectra. Note that (opposed to all other tested methods) B3-LYP does not reproduce the correct order of states for naphthalene. All methods underestimate the splitting. However, the error range is usually smaller than for the individual states, indicating at least some error compensation. DHDFs yield here the smallest errors, closely followed by LRC- ω PBE. The other methods show larger deviations, and global hybrids perform worst.

The results were further analyzed statistically. For all three energies analyzed in Figure 1, the mean (MDs) and mean absolute deviations from experiment (MADs) were calculated

Table 1. Mean (MD) and Mean Absolute Deviations (MAD) for Vertical Gas-Phase Excitation Energies into the 1L_a and 1L_b States for Five Linear Acenes (from naphthalene to hexacene) ^a

method		$\Delta E(^1L_a)$	$\Delta E(^1L_b)$	$\Delta E(^1L_a) - \Delta E(^1L_b)$
B2-PLYP	MD	-0.24	0.07	-0.31
	MAD	0.24	0.08	0.31
B2GP-PLYP	MD	-0.13	0.09	-0.22
	MAD	0.14	0.09	0.22
CC2 ^b	MD	0.05	0.22	-0.17
	MAD	0.08	0.22	0.17
LRC- ω PBE ^c	MD	0.01	0.33	-0.31
	MAD	0.04	0.33	0.31
LRC- ω PBEh ^c	MD	-0.08	0.35	-0.43
	MAD	0.08	0.35	0.43
LRC- μ BLYP ^{c,d}	MD	-0.3	0.18	-0.49
	MAD	0.3	0.18	0.49
LRC- μ BLYP ^{c,e}	MD	-0.01	0.31	-0.32
	MAD	0.04	0.31	0.32
B3-LYP ^c	MD	-0.45	0.15	-0.60 ^f
	MAD	0.45	0.15	0.60
BH-LYP	MD	-0.15	0.40	-0.56
	MAD	0.15	0.40	0.56
B2-LYP	MD	-0.13	0.42	-0.55
	MAD	0.13	0.42	0.55
B2GP-LYP	MD	-0.06	0.49	-0.54
	MAD	0.08	0.49	0.54
DFT/MRCI ^g	MD	-0.11	-0.09	-0.02
	MAD	0.11	0.10	0.03

^a Also the MDs and MADs for the state splittings [$\Delta E(^1L_a) - \Delta E(^1L_b)$] are given. All results are based on calculations with the cc-pVTZ basis. Theoretically corrected, experimental reference values were taken from ref 2. ^b Values taken from ref 2 ^c Values taken from ref 16 ^d With a long-range correction parameter of $\mu = 0.17$ a₀⁻¹. ^e With a long-range correction parameter of $\mu = 0.30$ a₀⁻¹. ^f The order of the two states is wrong for naphthalene. ^g On the basis of calculations with the TZVP basis; values taken from ref 12.

(see Table 1). Additionally, results for the BH-LYP functional (this work), for CC2 (taken from ref 2), for DFT/MRCI (taken from ref 12), and for LRC- μ BLYP with two different values for the long-range correction parameter μ (taken from ref 16) are also shown. The values reflect the conclusions drawn above. Global hybrid functionals benefit from higher amounts of Fock exchange regarding the 1L_a states (MAD = 0.45 eV for B3-LYP and MAD = 0.08 eV for B2GP-LYP), but the results become worse for the 1L_b states (MAD = 0.15 eV for B3-LYP and MAD = 0.49 eV for B2GP-LYP). The MADs for the state splittings range between 0.60 and 0.54 eV for the global hybrids. The results for the LRC functionals indicate the influence of the range-separation parameter. This can be particularly seen for the LRC- μ BLYP method, which gives reasonable 1L_b but only modest 1L_a excitation energies for a smaller μ , but shows worse (1L_b) or better agreement (1L_a) with a larger value. The remaining two LRC functionals reproduce 1L_a excitations very well; 1L_b states are similarly described as by global hybrids, though. DHDFs yield similar or even better results than the LRC functionals. B2GP-PLYP yields improved values compared to B2-PLYP, which is in accordance with previous benchmarks.^{9,23} Its MAD for the 1L_a

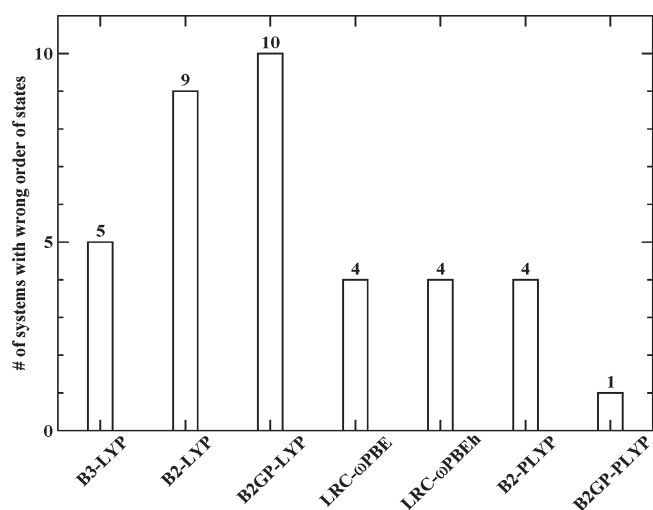


Figure 2. The number of polycyclic aromatic hydrocarbons (PAHs), for which the order of the 1L_a and 1L_b states is wrong. In total, 12 PAHs were tested. The results for B3-LYP and the two LRC methods, which were used for this analysis, were taken from ref 16.

states is slightly higher than for CC2 and the LRC functionals (MAD = 0.14 eV), but at the same time it can also describe the 1L_b states very well with an MAD of 0.09 eV, which is better than for CC2. B2-PLYP yields a higher MAD for 1L_a transitions (0.24 eV), but also a very low one for 1L_b (0.08 eV). DHDFs come closest to the very excellent results of the DFT/MRCI method, whose statistical values lie already within the estimated accuracy of the reference.

3.2. Nonlinear PAHs. Finally, the DHDFs are examined for nonlinear PAHs and compared to the B3-LYP, LRC- ω PBE, and LRC- ω PBEh functionals (results for those are again taken from ref 16). We took the investigated structures from the test set by Richard and Herbert. This set originally comprised 15 systems; however, not for every system were reference values available for both types of states. As we also always wanted to evaluate the state splittings and the energetic order of the states, we only took those 12 systems, for which all data were available (see Tables S3 and S4, Supporting Information for details on the systems). Reference values are based on measured absorption band maxima in solution;³⁵ however, as Richard and Herbert argued, those values should suffice to give a reliable estimate on a method's accuracy. Figure 2 depicts how many times a functional cannot reproduce the order of the two states correctly. For global hybrids with large amounts of nonlocal exchange, the results are unacceptable. B2-LYP predicts a wrong order in nine cases and B2GP-LYP in ten cases. Note, that it is expected that the BH-LYP functional, which is also popular in excited states applications, would not do a good job, either. As shown in Table 1 and also discussed in ref 23, BH-LYP results are practically identical to B2-LYP ones, due to similar amounts of Fock exchange. B3-LYP predicts a wrong order in five cases. The two LRC hybrids perform much better than the global ones and give wrong orders in only four cases. B2-PLYP performs similarly to the LRC methods (four cases). The best functional is B2GP-PLYP, which only gives the wrong order for one system (benzo[*b*]chrysene).

Richard and Herbert observed that whenever the B3-LYP functional performed well for the 1L_a states, the long-range corrected approaches yielded worse results and vice versa. Figure 3 shows similar plots like in ref 16. Part a shows the

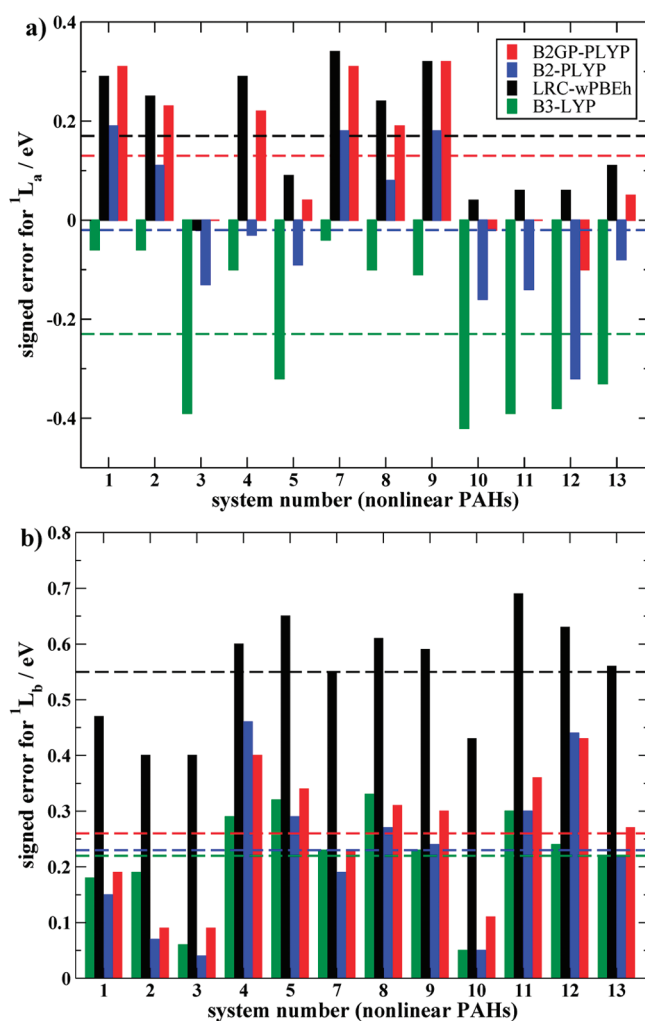


Figure 3. Signed errors (difference between theory and experiment) for various functionals for the 1L_a states (a) and the 1L_b states (b) for 12 nonlinear PAHs. All results are based on calculations with the cc-pVTZ basis. The results for B3-LYP and the LRC method were taken from ref 16. Experimental reference values were taken from ref 35. The numbers on the *x*-axes refer to the numbers given in ref 16; a full list of those systems is given in the Tables S3 and S4, Supporting Information.

signed errors for all tested nonlinear PAHs for B3-LYP, LRC- ω PBEh, and the two double-hybrids and compares them with their respective MDs. Like for the linear acenes, B3-LYP constantly underestimates the excitation energies, whereas LRC- ω PBEh tends to overestimate them; the DHDFs show a more balanced behavior and sometimes yield blue- or red-shifted energies, with usually smaller absolute errors. One clearly observes the behavior mentioned by Richard and Herbert when comparing B3-LYP to LRC- ω PBEh. Such a clear correlation between B3-LYP and the DHDFs is not seen; e.g., when the B3-LYP error is in an absolute range of 0.1 eV or less, B2-PLYP errors are often of the same order of magnitude. At first glance, there seems to be a correlation between the errors for LRC- ω PBEh and B2GP-PLYP. When sometimes the latter gives worse results than B2-PLYP, also the range-separated functional is worse and vice versa. However, there are also systems for which this is not the case (e.g., systems 5 and 13; see the Supporting Information for a description of these systems). Figure 3b shows the same analysis for the 1L_b states. All functionals overestimate

Table 2. Mean (MD) and Mean Absolute Deviations (MAD) for Vertical Excitation Energies into the 1L_a and 1L_b States of 12 Polycyclic Aromatic Hydrocarbons^a

method		$\Delta E(^1L_a)$	$\Delta E(^1L_b)$	$\Delta E(^1L_a) - \Delta E(^1L_b)$
B2-PLYP	MD	-0.02	0.23	-0.24 ^b
	MAD	0.14	0.23	0.26
B2GP-PLYP	MD	0.13	0.26	-0.13 ^c
	MAD	0.15	0.26	0.19
LRC- ω PBE ^d	MD	0.28	0.58	-0.30 ^b
	MAD	0.28	0.58	0.30
LRC- ω PBEh ^d	MD	0.17	0.55	-0.38 ^b
	MAD	0.18	0.55	0.38
B3-LYP ^d	MD	-0.23	0.22	-0.45 ^e
	MAD	0.23	0.22	0.45
B2-LYP	MD	0.13	0.65	-0.52 ^f
	MAD	0.16	0.65	0.52
B2GP-LYP	MD	0.20	0.77	-0.57 ^g
	MAD	0.21	0.77	0.57

^a Also the MDs and MADs for the state splittings [$\Delta E(^1L_a) - \Delta E(^1L_b)$] are given. All results are based on calculations with the cc-pVTZ basis. Experimental reference values were taken from ref 35. ^b The order of the two states is wrong in four cases. ^c The order of the two states is wrong in one case. ^d Values taken from ref 16. ^e The order of the two states is wrong in five cases. ^f The order of the two states is wrong in nine cases. ^g The order of the two states is wrong in 10 cases.

the excitation energies, with LRC- ω PBEh yielding the highest errors. Compared to the 1L_a states, the two DHDFs are more alike in this case and have overall the smallest errors.

The MDs and additionally the MADs are also shown in Table 2 (see Tables S3 and S4 of the Supporting Information for all calculated excitation energies). In all cases, the DHDFs show the best results. The MADs for the 1L_a transitions are 0.14 eV (B2-PLYP) and 0.15 eV (B2GP-PLYP). The MADs for the 1L_b transitions are more than halved compared to the LRC functionals (0.23 eV for B2-PLYP and 0.26 eV for B2GP-PLYP). Also the energy difference between the two states is better described by DHDFs. Considering the uncertainties due to, for example, solvent effects, it is expected that the errors would even be smaller, when considering a possible red-shift of about 0.1 eV to the calculated results (as also already noted in ref 16).

Note, that DHDFs do not contain any long-range corrections. It was shown that the description of CT transitions is better than for global hybrids but that they are clearly outmatched by range-separated methods for “true” CT states.⁹ On the other hand, DHDFs are not problematic for the 1L_a states, which is a further indication that they seem not to have large CT character.

4. SUMMARY

In this study, we have assessed the TD-B2-PLYP and TD-B2GP-PLYP methods on five linear and 12 nonlinear polycyclic aromatic hydrocarbons. We examined the errors for the two lowest lying states and also checked whether their energetic orders and differences were reproduced correctly. The two double-hybrids outmatched global hybrid and long-range corrected functionals and also the CC2 method in terms of accuracy and balanced distribution of the errors. This means that both 1L_a and 1L_b states are described on an equal footing, although their electronic character is rather different. Particularly, the B2GP-

PLYP functional with 65% Fock exchange and 36% nonlocal perturbative correlation yields very good results and also reproduces the order of the states best. Our findings are in accordance with previous benchmarks and good results found in a few recent applications.^{9,20,23,24,36} The advantage of the time-dependent double-hybrid methodology is the equally good description of different types of systems [for ground (see e.g. ref 37) and excited states] without an additional, system-dependent adjustment of parameters. The DHDFs perform better than exchange-only hybrids because they avoid self-interaction (overdelocalization) related errors by inclusion of a very high amount of Fock exchange and account for higher excitations by the nonlocal [i.e., (D)-type] correction for the excited state. We therefore again recommend double-hybrids for future applications to excited state problems but note that the theory in its present form is limited to low-lying states due to its perturbative character.

■ ASSOCIATED CONTENT

Supporting Information. All excitation energies for all systems and methods evaluated. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: grimmes@uni-muenster.de. Phone: (+49)-251-8336512.

■ ACKNOWLEDGMENT

L.G. was supported with a postdoctoral scholarship by the “Research School NRW—Molecules and Materials—A Common Design Principle” within the framework of the “NRW Graduate School of Chemistry”. We thank C. Mück-Lichtenfeld for his technical assistance.

■ REFERENCES

- Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 484–496.
- Grimme, S.; Parac, M. *Chem. Phys. Chem.* **2003**, *4*, 292–295.
- Parac, M.; Grimme, S. *Chem. Phys.* **2003**, *292*, 11–21.
- Casida, M. E. Time-Dependent Density Functional Response Theory for Molecules. In *Recent Advances in Density Functional Methods*; D. P. Chong, Ed.; World Scientific: Singapore, 1995; Vol. 1, pp 155–192.
- Gross, E. K. U.; Dobson, J. F.; Petersilka, M. *Top. Curr. Chem.* **1996**, *181*, 81–172.
- Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454–464.
- Time-Dependent Density Functional Theory, Lecture Notes Physics* 706; Marques, M. A. L., Ulrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K. U., Eds.; Springer: Berlin-Heidelberg, 2006.
- Jacquemin, D.; Wathélet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.
- Goerigk, L.; Grimme, S. *J. Chem. Phys.* **2010**, *132*, 184103.
- Dierksen, M.; Grimme, S. *J. Phys. Chem. A* **2004**, *108*, 10225–10237.
- Grimme, S.; Waletzke, M. *J. Chem. Phys.* **1999**, *111*, 5645–5655.
- Marian, C.; Gilka, N. *J. Chem. Theory Comput.* **2008**, *4*, 1501–1515.
- Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135.
- Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Theor. Chem. Acc.* **2011**, *128*, 127–136.

- (15) Wong, B. M.; Hsieh, T. H. *J. Chem. Theory Comput.* **2010**, *6*, 3704–3712.
- (16) Richard, R. M.; Herbert, J. M. *J. Chem. Theory Comput.* **2011**, *7*, 1296–1306.
- (17) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818–2820.
- (18) Stein, T.; Kronik, L.; Baer, R. *J. Chem. Phys.* **2009**, *131*, 244119.
- (19) Kuritz, N.; Stein, T.; Kronik, R. B. L. *J. Chem. Theory Comput.* **2011**, *7*, 2408–2415.
- (20) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.
- (21) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (22) Karton, A.; Tarnopolsky, A.; Lamère, J. F.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.
- (23) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611–4620.
- (24) Goerigk, L.; Grimme, S. *J. Phys. Chem. A* **2009**, *113*, 767–776.
- (25) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. *Chem. Phys. Lett.* **1994**, *219*, 21–29.
- (26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (27) Ahlrichs, R.; et al. *TURBOMOLE, version 5.9*; Universität Karlsruhe, 2008. See <http://www.turbomole.com>.
- (28) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (29) Häser, M.; Ahlrichs, R. *J. Comput. Chem.* **1989**, *10*, 104–111.
- (30) Furche, F.; Rappoport, D. In *Theoretical and Computational Chemistry*; Olivucci, M., Ed.; Elsevier: Amsterdam, 2005; Vol. 16
- (31) Grimme, S. *RICC: A coupled-cluster program using the RI approximation*; University of Münster, 2007.
- (32) Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97*, 331–340.
- (33) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (34) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.
- (35) Birks, J. B. *Photophysics of Aromatic Molecules*; Wiley: London, 1970.
- (36) Vintonyak, V.; Warburg, K.; Kruse, H.; Grimme, S.; Huebel, K.; Rauh, D.; Waldmann, H. *Angew. Chem., Int. Ed.* **2010**, *49*, S902–S905.
- (37) Goerigk, L.; Grimme, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

Scalar Relativistic Computations of Nuclear Magnetic Shielding and g -Shifts with the Zeroth-Order Regular Approximation and Range-Separated Hybrid Density Functionals

Fredy Aquino,[†] Niranjan Govind,[‡] and Jochen Autschbach^{*,†}

[†]Department of Chemistry, State University of New York at Buffalo, Buffalo, New York 14260-3000

[‡]William R. Wiley Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 902 Battelle Blvd, P.O. Box 999, Mail Stop K8-91 Richland, Washington 99352, United States

 Supporting Information

ABSTRACT: Density functional theory (DFT) calculations of NMR chemical shifts and molecular g tensors with Gaussian-type orbitals are implemented via second-order energy derivatives within the scalar relativistic zeroth order regular approximation (ZORA) framework. Nonhybrid functionals, standard (global) hybrids, and range-separated (Coulomb-attenuated, long-range corrected) hybrid functionals are tested. Origin invariance of the results is ensured by use of gauge-including atomic orbital (GIAO) basis functions. The new implementation in the NWChem quantum chemistry package is verified by calculations of nuclear shielding constants for the heavy atoms in HX ($X = \text{F, Cl, Br, I, At}$) and H_2X ($X = \text{O, S, Se, Te, Po}$) and ^{125}Te chemical shifts in a number of tellurium compounds. The basis set and functional dependence of g -shifts is investigated for 14 radicals with light and heavy atoms. The problem of accurately predicting ^{19}F NMR shielding in $\text{UF}_{6-n}\text{Cl}_n$, $n = 1-6$, is revisited. The results are sensitive to approximations in the density functionals, indicating a delicate balance of DFT self-interaction vs correlation. For the uranium halides, the range-separated functionals are not clearly superior to global hybrids.

1. INTRODUCTION

Magnetic resonance spectroscopy is of high importance in basic and applied research. In the past decades, much effort has been directed at computations of such parameters for molecules starting from first principles theory.¹⁻⁷ This report is concerned with NMR nuclear magnetic shielding as well as electron paramagnetic resonance (EPR) g -shifts and, more specifically, with relativistic effects. A free electron has a g factor of $g_e \approx 2.0023$. A deviation of g from g_e , associated with the effective spin of a molecule, is linked to spin-orbit (SO) coupling, which is a relativistic effect. The g shift tensor Δg quantifies the deviation from the isotropic free electron case and represents one of the sets of parameters that defines the EPR spectrum.⁸ The g tensor is also important when considering paramagnetic effects in the NMR of open shell molecules.⁹⁻¹² The “regular” NMR shielding has a nonvanishing, nonrelativistic limit. However, due to the nature of the quantum mechanical operators involved in its calculation, relativistic effects on NMR shieldings can be highly significant and have been the subject of much theoretical work spanning several decades.^{2,13-19}

The method of choice for first-principles computations of magnetic resonance parameters of large molecules with heavy elements and metal complexes is density functional theory (DFT), due to its attractive balance of computational cost and the accuracy level that can be attained in the computations. The combination of DFT with the approximate two-component relativistic zeroth-order regular approximation (ZORA)^{20,21} represents an efficient way of carrying out correlated relativistic electronic structure calculations. ZORA is one among several efficient methods available to perform relativistic electronic structure

and magnetic property calculations. Among the more widely applied methods, we mention the two-component Douglas-Kroll-Hess approach²²⁻²⁴ and (usually somewhat more demanding) fully relativistic methods.²⁵ The ZORA Hamiltonian has been found to be a suitable choice for computations of magnetic resonance parameters²⁶⁻³⁵ because it can accurately generate relativistic effects for molecular properties that are dominated by valence orbital contributions, such as NMR chemical shifts, J -coupling, and g -shifts. Further, it is straightforward to derive ZORA perturbation operators for magnetic properties, and to calculate matrix elements thereof, using numerical integration. In molecular DFT calculations where numerical integration of the exchange-correlation (XC) potential and the associated response kernels is standard practice, one might argue that there is not much to gain from calculating ZORA perturbation operator matrices fully analytically because such techniques require additional approximations.³⁶ We note, however, that analytic integrals can be utilized in ZORA computations, for instance by separating nonrelativistic contributions from integrals involving relativistic operators.³⁷ Magnetic resonance calculations reported in the works cited above have utilized the ZORA implementations of the Amsterdam Density Functional (ADF) package.³⁸ This code employs Slater-type orbital (STO) basis sets and density fitting for calculating the Coulomb potential and has already for a long time incorporated methods for ZORA calculations of a variety of NMR and EPR parameters as well as other molecular response properties. Hybrid DFT functionality

Received: June 15, 2011

Published: September 14, 2011

for magnetic resonance parameters in this STO-based program has been reported recently.^{28,31,35,39}

Herein, we report the development of a ZORA methodology for scalar relativistic DFT computations for NMR shielding tensors and EPR Δg tensors in the NWChem package,³⁷ utilizing gauge-including atomic orbitals (GIAOs) to ensure origin-invariant results. The implementation makes use of the coupled-perturbed Hartree–Fock equation solver originating in Dupuis' nonrelativistic GIAO shielding implementation in NWChem,⁴⁰ which would have to be redesigned to enable spin–orbit calculations. NWChem employs Gaussian-type orbital (GTO) basis sets for molecular calculations. Its DFT module has recently been extended^{41–46} to allow for computations with range-separated hybrid functionals.^{47–49} This class of functionals shows promise for alleviating certain problems in DFT and time-dependent DFT calculations of response properties such as the correct asymptotic behavior of the potential and the treatment of charge transfer. The ZORA implementation for NMR and EPR utilizes numerical integration methods for operator matrix elements that we recently implemented in NWChem for relativistic calculations of electric field gradients.⁵⁰ The availability of the new static linear response ZORA functionality developed for this work is used to investigate the performance of relativistic magnetic resonance DFT calculations in conjunction with range-separated hybrid functionals. Further, we aim to test the suitability of relatively common Gaussian-type basis sets (such as the ones used in previous work on ZORA electric field gradients) for scalar ZORA NMR chemical shift and EPR g shift computations.

There is a body of data showing that nonhybrid functionals occasionally perform poorly in calculations of g -shifts ranging from light organic radicals to heavy metal complexes^{10,24,51–58} and NMR shielding in metal complexes.^{13,19,59–61} Standard (global) hybrid functionals also do not consistently perform well, in particular for metal complexes. It is therefore interesting to investigate the performance of range-separated hybrids in conjunction with a relativistic method. As an example, fluorine chemical shifts in U(VI) complexes have been determined previously with nonrelativistic hybrid DFT computations employing a relativistic ECP for uranium.^{60,62} It was found that the ^{19}F chemical shifts are sensitive to the computational model. For instance, the ordering of the fluorine shift in UF_6 versus *fac*- UF_3Cl_3 is predicted incorrectly with nonhybrid functionals and the B3LYP hybrid functional, whereas the correct ordering is obtained with the BHLYP functional, which has a much larger fraction of Hartree–Fock (HF) exchange than B3LYP (50 vs 20%). In a benchmark study of molecular g -tensors, we determined recently that the PBE0 hybrid performed similarly to the nonhybrid PBE functional across a small-molecule test set including inorganic and organic radicals with light and heavy metal atoms,³⁵ but it is unclear yet how all-electron relativistic calculations with range-separated hybrid functionals perform in comparison. For the case of the ^{19}F shielding in $\text{UF}_{6-n}\text{Cl}_n$, $n = 1–6$, it is shown herein that there is a difficult balance between reproducing the magnitude of the shielding and trends such as UF_6 versus *fac*- UF_3Cl_3 where hybrid functionals with a large fraction of HF exchange perform better, and other trends among the set of complexes that appear to be better reproduced with nonhybrid functionals and B3LYP, albeit with an overall too small magnitude of the ^{19}F shieldings. The range-separated hybrids yield shielding constants that are closer to experimental results and to calculations with 50% global hybrids, without achieving a breakthrough improvement over the latter. For the

g shift test set, we find comparable results calculated with CAM-B3LYP and the PBE0 global hybrid (25% HF exchange).

Following this Introduction, theoretical details of the ZORA linear response shielding and g shift calculations are provided in section 2 along with details about the implementation, such as the numerical integration of the relevant perturbation operator matrix elements. Computational details are provided in section 3. Results from various benchmark computations are provided in section 4 in order to verify the implementations and investigate the performance of nonhybrid vs hybrid DFT with global fixed fractions of HF exchange vs range-separated hybrid functionals. This work concludes with a brief summary and outlook in section 5.

2. THEORETICAL METHODS

A component of the chemical shift tensor can be calculated within a DFT framework from variational perturbation theory as a second derivative of the energy E as

$$\sigma_{uv} = \frac{\partial^2 E}{\partial B_u \partial \mu_v} \quad (1)$$

Here, B_u is the component of an external magnetic field, and μ_v is a component of the nuclear spin magnetic moment vector of the nucleus at which the shielding tensor is calculated, and $u, v \in \{x, y, z\}$. Here and elsewhere, it is assumed that magnetic-field-related perturbation parameter derivatives are taken at $B_u = 0$ and $\mu_v = 0$. Within a scalar relativistic or nonrelativistic framework, one can also define the components of the electronic g -tensor as a second derivative of the molecular energy⁶³

$$g_{uv} = \frac{1}{\beta_e} \frac{\partial^2 E}{\partial B_u \partial S_v} \quad (2)$$

in which case spin–orbit coupling is treated as a perturbation and not included in the ground state calculation. Here, S_v is a component of the effective spin vector of the molecule. The Bohr magneton $\beta_e = e\hbar/(2m_e)$ enters the expression based on usual conventions for the EPR spin Hamiltonian.⁸ In the Dirac theory of the electron, the free electron g_e is exactly 2. In the following, atomic units where $g_e\beta_e = 1$ are used. The calculations determine g -shifts Δg directly as deviations from the free electron g value. Therefore, it is not necessary to specify a particular value for g_e in the implementation.

As Schreckenbach and Ziegler have demonstrated within the framework of the Pauli Hamiltonian,⁶³ the computational machineries for shielding and Δg tensors based on eqs 1 and 2 share many common components. Functionality for shielding and Δg tensors can therefore be developed in a concerted fashion. The calculations for Δg and NMR shielding reported here include scalar relativistic effects variationally in E . For g -shifts, spin–orbit (SO) coupling is treated as a perturbation. For shielding tensors, the effects from SO coupling are presently neglected. Some research groups have developed theoretical methods for calculations of g tensors as *first* derivatives of the energy instead, with SO coupling included variationally in the unperturbed set of orbitals,^{34,55–57,64} We plan to investigate such an approach in conjunction with ZORA in a follow-up report. A recent comparison of the first-order and second-order derivative methods in their respective implementations in the ADF code using ZORA in conjunction with STO basis sets has shown that they can yield comparable results³⁵ overall, with the notable exception of the

known symmetry-related breakdown of the second-order approach for the Δg tensor components in linear molecules.⁶⁵ In this context, we note recent work by Hrobarik et al. where differences in treating SO coupling as a first-order perturbation versus as a higher order one were found to be significant.⁶⁶

For the shielding tensor calculations, a closed-shell reference is assumed. The g -tensor calculations are based on spin-unrestricted DFT computations. In a scalar relativistic spin-unrestricted Kohn–Sham DFT approach with singly occupied pure α and β spin orbitals, the maximum spin projection is

$$S_z = \frac{n_\alpha - n_\beta}{2} \quad (3)$$

where n_α and n_β are the numbers of occupied orbitals of α and β spins, respectively. The “spin derivative” in eq 2 as well as the associated operator derivatives (vide infra) are understood to be taken in the expectation value after the action of spin dependent operators on the orbitals has been considered. This procedure leads to an overall factor of S_z^{-1} in the final expression for the g -tensor and to contributions of opposite signs from α and β spin orbitals, respectively. For details and a review of pertinent literature see refs 5, 6, and 35 as well as the review articles by Kaupp et al. and Neese cited in the Introduction. In the NMR shielding tensor calculations, the contributions from α and β orbitals reinforce each other.

The ZORA one-electron Fock operator used in DFT with a local effective potential V reads in atomic units:²⁰

$$\begin{aligned} \hat{h} &= V + \frac{1}{2} (\vec{\sigma} \cdot \hat{\mathbf{p}}) \mathcal{K} (\vec{\sigma} \cdot \hat{\mathbf{p}}) \\ &= V + \frac{1}{2} \hat{\mathbf{p}} \cdot \mathcal{K} \hat{\mathbf{p}} + \frac{1}{2} i \vec{\sigma} \cdot (\hat{\mathbf{p}} \mathcal{K} \times \hat{\mathbf{p}}) \end{aligned} \quad (4)$$

with

$$\mathcal{K} = \frac{2c^2}{2c^2 - V} \quad (5)$$

Further, $\vec{\sigma}$ is the 3-vector of 2×2 Pauli spin matrices, with components $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$, and $\hat{\mathbf{p}} = -i\nabla$ is the momentum operator. The last term in eq 4 is the ZORA spin–orbit operator which, for the purpose of this work, is neglected in the shielding computations and included as a perturbation in the g shift calculations. The unperturbed Kohn–Sham orbitals $\varphi_i^{(0)}$ are determined with the scalar relativistic ZORA (ZORA-SR) Fock operator

$$\hat{h}^{(0)} = V + \frac{1}{2} \hat{\mathbf{p}} \cdot \mathcal{K} \hat{\mathbf{p}} \quad (6)$$

The potential V in eqs 4 and 6 is determined self-consistently for systems with more than one (spin) orbital and may be adapted in the usual way for standard and range-separated hybrid functionals to include exact exchange contributions. The potential V used to construct \mathcal{K} in eq 5 is in the NWChem implementation approximated as a sum of (local) atomic Hartree potentials V_A as $V \approx \sum_A V_A$.^{37,50} A related approach was taken by van Wüllen for a ZORA implementation in the Turbomole code⁶⁷ and by Philipsen et al. for an implementation in the Amsterdam Density functional (ADF) package.⁶⁸ Such approximations in the ZORA operator have since been demonstrated to be a practical and accurate way to enable ZORA relativistic computations for large systems.

The shielding and the g -tensor involve magnetic field derivatives. In order to avoid origin-dependent results, we adopt a “gauge-including atomic orbital” (GIAO)^{69,70} basis set. In terms of a standard atom-centered GTO or STO basis set $\{\chi_s\}$, the GIAOs $\xi_s(\mathbf{B})$ are given as

$$\xi_s(\mathbf{B}) = \chi_s \exp \left[-\frac{i}{2} (\mathbf{B} \times \mathbf{R}_s) \cdot \mathbf{r} \right] \quad (7)$$

where \mathbf{R}_s is the center of the AO basis function χ_s . Below, the electron location with respect to a basis function center, $\mathbf{r}_s = \mathbf{r} - \mathbf{R}_s$, is used frequently. The coefficients for the unperturbed MOs in the basis set are denoted as $C_{ri}^{(0)}$. The basis set coefficients of the magnetic field perturbed MOs for field direction u are $C_{ri}^{(u)}$. It is beneficial to utilize elements of the scalar relativistic density (+) and spin-density (−) matrices in the AO basis given by

$$P_{rs}^{(0)\alpha \pm \beta} = \sum_i n_i^\alpha C_{ri}^{(0)\alpha} C_{si}^{*(0)\alpha} \pm \sum_i n_i^\beta C_{ri}^{(0)\beta} C_{si}^{*(0)\beta} \quad (8a)$$

and

$$\begin{aligned} P_{rs}^{(u)\alpha \pm \beta} &= \sum_i n_i^\alpha [C_{ri}^{(0)\alpha} C_{si}^{*(u)\alpha} + C_{ri}^{(u)\alpha} C_{si}^{*(0)\alpha}] \\ &\pm \sum_i n_i^\beta [C_{ri}^{(0)\beta} C_{si}^{*(u)\beta} + C_{ri}^{(u)\beta} C_{si}^{*(0)\beta}] \end{aligned} \quad (8b)$$

where the n_i^γ , $\gamma = \alpha$ and β , are the occupation numbers for the spin orbitals. For the magnetic field perturbation and real one-component unperturbed orbitals, the perturbed orbital and spin-density matrices are imaginary. For the calculation of the perturbed MO coefficients, we employ a modified version of the coupled-perturbed Kohn–Sham (CPKS) equation solver that was originally implemented in NWChem by Dupuis⁴⁰ for the purpose of nonrelativistic NMR calculations. The ZORA functionality incorporates extensions for the computation of DFT contributions to the perturbed Fock matrices used to determine the $C_{ri}^{(u)}$. These modifications entail the use of numerical integration to determine AO matrix elements for a scalar ZORA external magnetic field perturbation operator suitable for GIAO computations:

$$\hat{h}_{rs}^{(u)} = -\frac{i}{4} [\mathcal{K} (\mathbf{r}_r \times \nabla)_u + (\mathbf{r}_s \times \nabla)_u \mathcal{K}] \quad (9)$$

and the use of the unperturbed ZORA Fock operator, eq 6, in GIAO magnetic field derivatives of the Fock operator matrix elements needed in the CPKS solver,⁴⁰ which requires matrix elements of

$$\hat{h}_{rs}^{\text{SR},(u)} = -\frac{i}{4} \nabla \mathcal{K} [\mathbf{r} \times (\mathbf{R}_s - \mathbf{R}_r)]_u \cdot \nabla \quad (10)$$

Another modification has been necessary in the CPKS solver related to the calculation of two-electron integrals and their GIAO magnetic field derivatives, which are calculated in NWChem using Rys quadrature. These modifications are described in the Appendix. For further details regarding the CPKS procedure, we refer the reader to Dupuis’ original paper,⁴⁰ which expands on many of the associated technical details.

Working expressions that were implemented for the calculation of GIAO chemical shift and Δg tensor elements are as follows:

Diamagnetic nuclear magnetic shielding, nucleus Q:

$$\sigma_{uv}^d = \sum_{r,s} P_{rs}^{(0)\alpha + \beta} \left[\langle \chi_r | \frac{1}{2} [\hat{h}_{r,Q}^{(u,v)} + \hat{h}_{s,Q}^{(u,v)}] | \chi_s \rangle + \langle \chi_r | \hat{R}_{rs}^{(u,v)} | \chi_s \rangle \right] \quad (11a)$$

Paramagnetic nuclear magnetic shielding, nucleus Q:

$$\sigma_{uv}^p = \sum_{r,s} \left[2\text{Re} \left\{ P_{rs}^{(u)\alpha + \beta} \langle \chi_r | \hat{h}_{s,Q}^{(v)} | \chi_s \rangle \right\} + P_{rs}^{(0)\alpha + \beta} \frac{i}{2} (\mathbf{R}_r \times \mathbf{R}_s)_u \langle \chi_r | \hat{h}_{s,Q}^{(v)} | \chi_s \rangle \right] \quad (11b)$$

Diamagnetic Δg :

$$\Delta g_{uv}^d = \frac{4}{n_\alpha - n_\beta} \sum_{r,s} P_{rs}^{(0)\alpha - \beta} [\langle \chi_r | \hat{h}_s^{(u,v)} | \chi_s \rangle + \langle \chi_r | \hat{R}_{rs}^{(u,v)} | \chi_s \rangle] \quad (12a)$$

Paramagnetic Δg :

$$\Delta g_{uv}^p = \frac{4}{n_\alpha - n_\beta} \sum_{r,s} \left[2\text{Re} \left\{ P_{rs}^{(u)\alpha - \beta} \langle \chi_r | \hat{h}_s^{(v)} | \chi_s \rangle \right\} + P_{rs}^{(0)\alpha - \beta} \frac{i}{2} (\mathbf{R}_r \times \mathbf{R}_s)_u \langle \chi_r | \hat{h}_s^{(v)} | \chi_s \rangle \right] \quad (12b)$$

The prefactors in eq 12 represent $1/(\beta_e S_z)$, where eq 3 has been used for the effective spin. Going back to a suggestion by Fukui,⁷¹ GIAO-related terms have been grouped to yield diamagnetic and paramagnetic shielding and Δg tensors that are individually origin-independent. The total tensors are given by the sum of the paramagnetic and diamagnetic components. The various operators in eqs 11 and 12 are as follows:

Shielding tensor, eq 11:

$$\hat{h}_Q^{(v)} = -\frac{i}{2} \left[\mathcal{H} \frac{(\mathbf{r}_Q \times \nabla)_v}{r_Q^3} + \frac{(\mathbf{r}_Q \times \nabla)_v}{r_Q^3} \mathcal{H} \right] \quad (13a)$$

$$\hat{h}_{s,Q}^{(u,v)} = \frac{\mathcal{H}}{2r_Q^3} (\delta_{uv} \mathbf{r}_Q \cdot \mathbf{r}_s - r_{Q,u} r_{s,v}) \quad (13b)$$

$$\hat{R}_{rs,Q}^{(u,v)} = \frac{1}{4} \left[\frac{(\mathbf{r}_Q \times \nabla)_v}{r_Q^3} \mathcal{H} [\mathbf{r}_r \times (\mathbf{R}_s - \mathbf{R}_r)]_u + \right. \quad (13c)$$

$$\left. \mathcal{H} [\mathbf{r}_s \times (\mathbf{R}_s - \mathbf{R}_r)]_u \frac{(\mathbf{r}_Q \times \nabla)_v}{r_Q^3} \right] \quad (13d)$$

Δg tensor, eq 12:

$$\hat{h}^{(v)} = -\frac{i}{2} (\nabla \mathcal{H} \times \nabla)_v \quad (13e)$$

$$\hat{h}_s^{(u,v)} = \frac{1}{4} \{ \delta_{uv} \nabla \cdot (\mathcal{H} - 1) \mathbf{r}_s - \nabla_u (\mathcal{H} - 1) r_{s,v} \} \quad (13f)$$

$$\hat{R}_{rs}^{(u,v)} = \frac{i}{2} [\mathbf{r}_s \times (\mathbf{R}_s - \mathbf{R}_r)]_u \hat{h}^{(v)} \quad (13g)$$

Curly brackets indicate where derivatives are taken only in the operator, i.e., where ∇ is not acting on any basis functions or orbitals to the right. Using $(\mathcal{H} - 1)$ instead of \mathcal{H} in eq 13f has the effect of subtracting the free electron g value from the calculated g -tensor, and thus eq 12 yields Δg . Magnetic perturbation operators in ZORA typically afford derivatives of \mathcal{H} .

AO matrix elements of the operators are calculated in our implementation by numerical integration. By using the turnover rule for the momentum operator and/or partial integration, derivatives of \mathcal{H} can be switched over to the basis functions χ_r and χ_s instead.

Matrix elements for the external magnetic field operator, eq 9, are calculated as

$$H_{rs}^{(u)} = \frac{i}{4} (L_{rs,mm} - L_{sr,mm}^* - [L_{rs,nm} - L_{sr,nm}^*]) \quad (14a)$$

with $L_{rs,mm}$ given in terms of basic AO integrals as

$$L_{rs,mm} = \langle \frac{\partial \chi_r}{\partial r_n} | \mathcal{H} r_m | \chi_s \rangle - R_{s,m} \langle \frac{\partial \chi_r}{\partial r_n} | \mathcal{H} | \chi_s \rangle \quad (14b)$$

where $umn = xyz, yzx, zxy$. Matrix elements for the scalar relativistic operator of eq 10 were implemented as

$$H_{rs}^{\text{SR},(u)} = \frac{i}{4} [R_{sr,b} \sum_p Y_{rs,pp,a} - R_{sr,a} \sum_p Y_{rs,pp,b}] \quad (14c)$$

with $R_{sr} = \mathbf{R}_s - \mathbf{R}_r$, $uab = xyz, yzx, zxy$, and basic AO integrals $Y_{rs,pq,a}$ given as

$$Y_{rs,pq,a} = \langle \frac{\partial \chi_r}{\partial r_p} | r_a \mathcal{H} | \frac{\partial \chi_s}{\partial r_q} \rangle \quad (14d)$$

Matrix elements of the ZORA analog of the paramagnetic spin-orbital (PSO) operator, eq 13a, were implemented as

$$H_{rs,Q}^{(v)} = \frac{i}{2} (N_{rs,mm} - N_{sr,mm}^* - [N_{rs,nm} - N_{sr,nm}^*]) \quad (14e)$$

with $N_{rs,mm}$ given as

$$N_{rs,mm} = \langle \frac{\partial \chi_r}{\partial r_n} | \mathcal{H} \frac{r_{Q,m}}{r_Q^3} | \chi_s \rangle \quad (14f)$$

where $vmn = xyz, yzx, zxy$.

Matrix elements of the ZORA diamagnetic GIAO shielding operator eq 13b, were implemented as

$$H_{rs,Q}^{(u,v)} = \frac{1}{4} \begin{cases} -I_{rs,Q}^{(u,v)} & u \neq v \\ \sum_{p \neq u} I_{rs,Q}^{(p,p)} & u = v \end{cases} \quad (14g)$$

where $I_{rs,Q}^{(u,v)}$ is given in terms of basic AO integrals as

$$I_{rs,Q}^{(u,v)} = 2 \langle \chi_r | \frac{\mathcal{H} r_u r_v}{r_Q^3} | \chi_s \rangle + R_{Q,u} (R_{r,v} + R_{s,v}) \langle \chi_r | \frac{\mathcal{H}}{r_Q^3} | \chi_s \rangle - (R_{r,v} + R_{s,v}) \langle \chi_r | \frac{\mathcal{H} r_u}{r_Q^3} | \chi_s \rangle - 2R_{Q,u} \langle \chi_r | \frac{\mathcal{H} r_v}{r_Q^3} | \chi_s \rangle \quad (14h)$$

The matrix elements of eq 13c were implemented as follows:

$$R_{rs,Q}^{(u,v)} = \frac{1}{4} \{ R_{sr,b} [B_{rs,mm,a} - B_{sr,mm,a}^* - (B_{rs,nm,a} - B_{sr,nm,a}^*)] - R_{sr,a} [B_{rs,mm,b} - B_{sr,mm,b}^* - (B_{rs,nm,b} - B_{sr,nm,b}^*)] - \frac{i}{2} (\mathbf{R}_r \times \mathbf{R}_s)_u H_{rs,Q}^{(v)} \} \quad (14i)$$

where $uab = xyz, yzx, zxy$; $vmn = xyz, yzx, zxy$; and $B_{rs, mn, a}$ is given as

$$B_{rs, mn, a} = \langle \chi_r | r_a \mathcal{K} \frac{r_{Q, m}}{r_Q^3} \frac{\partial \chi_s}{\partial r_n} \rangle \quad (14j)$$

with $r_a = x, y, z$. Further, $\mathbf{R}_{sr} = \mathbf{R}_s - \mathbf{R}_r$.

Matrix elements for the spin-orbit operator derivative, eq 13e, needed for the Δg tensor are calculated as

$$H_{rs}^{(v)} = \frac{i}{2} \int d^3r (\mathcal{K} - 1) [\{\nabla \chi_r^*\} \times \{\nabla \chi_s\}] \quad (14k)$$

using AO integrals of the form $\langle (\partial \chi_r / \partial r_m) | \mathcal{K} - 1 | (\partial \chi_s / \partial r_n) \rangle$. In eq 14k, use is made of the fact that $\nabla \mathcal{K} \times \nabla = \nabla (\mathcal{K} - 1) \times \nabla$ in order to eliminate vanishing contributions.³⁵ For the bilinear perturbation operator, eq 13f, in the Δg tensor, matrix elements are calculated as

$$H_{rs}^{(u, v)} = -\frac{1}{4} [\delta_{uv} \sum_n (M_{rs, mn, s} + M_{sr, mn, s}^*) - (M_{rs, uv, s} + M_{sr, uv, s}^*)] \quad (14l)$$

with $M_{rs, uv, t}$ given as

$$M_{rs, uv, t} = \langle \frac{\partial \chi_r}{\partial r_u} | (\mathcal{K} - 1) r_{t, v} | \chi_s \rangle \quad (14m)$$

where $r_{t, v} = r_v - R_{t, v}$.

Finally, we have the matrix elements of eq 13g, which were implemented as

$$\begin{aligned} R_{rs}^{(u, v)} = & -\frac{1}{4} [R_{sr, b} [A_{rs, mn, a} - A_{sr, mn, a}^*] \\ & - R_{sr, a} [A_{rs, mn, b} - A_{sr, mn, b}^*] + \\ & \delta_{uv} \sum_n R_{sr, n} \langle \chi_r | (\mathcal{K} - 1) \frac{\partial \chi_s}{\partial r_n} \rangle - R_{sr, v} \langle \chi_r | (\mathcal{K} - 1) \frac{\partial \chi_s}{\partial r_u} \rangle \\ & - \frac{i}{2} (\mathbf{R}_r \times \mathbf{R}_s)_u H_{rs}^{(v)} \end{aligned} \quad (14n)$$

with

$$A_{rs, mn, a} = \langle \frac{\partial \chi_r}{\partial r_m} | r_a (\mathcal{K} - 1) \frac{\partial \chi_s}{\partial r_n} \rangle \quad (14o)$$

Equation sets 11, 12, 13, and 14 form the working expressions for the ZORA implementation of NMR shielding and Δg tensors in NWChem.

The formalism outlined here corresponds to “unscaled” ZORA computations. A “scaled” ZORA variant has been suggested by Wolff et al.²⁶ In the scaled ZORA approach, the shielding tensor is formulated in terms of a sum of derivatives of the Kohn–Sham orbitals φ_i and their energies ε_i . As was shown by van Lenthe,⁷² the ZORA orbital energies of one-electron systems can be improved considerably toward the fully relativistic value by applying a scaling factor

$$\varepsilon_i^{\text{scaled-ZORA}} = S_i \varepsilon_i^{\text{ZORA}} \quad (15)$$

where, in a scalar relativistic ZORA framework

$$S_i = \left[1 + \langle \varphi_i | \hat{\mathbf{p}} \frac{c^2}{[2c^2 - V]^2} \hat{\mathbf{p}} | \varphi_i \rangle \right]^{-1} \quad (16)$$

The scaling factors can be conveniently included in the unperturbed and perturbed density matrices in eqs 8a and 8b.⁵⁰ For valence orbitals, which dominate the chemical shifts, the scaling factors are close to 1 and do not severely alter the results. Larger effects are generally obtained for core orbitals where the unscaled ZORA orbital energies and related terms in derivative properties are not very accurate.

3. COMPUTATIONAL DETAILS

The new implementation and computations reported herein are based on a 2011 developer’s version of the open source NWChem package.^{73–75} Recently developed ZORA functionality reported by Nichols et al.³⁷ and by us⁵⁰ has been utilized and extended for scalar ZORA linear response DFT computations as described in section 2. Consistent with a previous ZORA implementation of g -shifts as second order derivatives in the Amsterdam Density Functional (ADF) package,^{38,35} spin–other-orbit (SOO) terms have been neglected since our main focus is on systems with heavy atoms. For a justification of this approximation, see ref 76. Finite nuclear models were not used for the present study.

Computations that were performed mainly to validate the NWChem ZORA-SR shielding module employed the set of diatomic molecules HX (X = F, Cl, Br, I, At) and the H₂X (X = O, S, Se, Te, Po) series. Geometries for HX and H₂X were obtained from refs 77 and 78, respectively. For the heavy atoms X, we employed a Gaussian-type atomic orbital (GTO) basis set with sufficient flexibility in the core region to be able to describe the scalar ZORA effects on the orbitals. The ANO-RCC basis sets by Roos et al.⁷⁹ in a fully uncontracted fashion were found to be suitable for this purpose (for a related ZORA benchmark of electric field gradients, see ref 50). For hydrogen, the TZVPP basis set by Weigend and Ahlrichs⁸⁰ was used. This and the other basis sets were downloaded from the EMSL basis set exchange.^{81,82} A second set of calculations aimed at verifying the implementation was performed for a set of selected molecules containing tellurium. These systems have been used previously to explore scalar relativistic effects on NMR chemical shifts.^{83–85} For the Te benchmark, we employed two sets of geometries: (a) geometries obtained from X-ray diffraction, as reported in ref 86 for TeCl₄, ref 87 for TeF₂C₂F₆, ref 88 for TeMe₂Cl₂, ref 89 for TeO₆H₆, ref 90 for TeF₆, ref 91 for TeMe₂, ref 92 for TeMe₄, and ref 93 for TeH₂; (b) optimized geometries obtained with ADF using the Becke88⁹⁴+Perdew86⁹⁵ (BP) functional and a triple- ζ doubly polarized ZORA STO basis set (TZ2P). For the NWChem NMR calculations, we employed the uncontracted ANO-RCC basis set for tellurium and the 6-311G* Pople GTO basis set on the ligands. For comparison, ZORA-SR NMR computations were carried out with a developers version (pre-2011 release) of ADF using a quadruple- ζ multiply polarized (QZ4P) ZORA STO basis set for all atoms. For the validation and comparison with ADF, the following functionals were applied in the NMR computations: BP and the Becke three-parameter Lee–Yang–Parr (B3LYP) hybrid.⁹⁶ The performance of the NWChem ZORA-SR shielding module for the evaluation of ¹²⁵Te chemical shifts has been further tested by comparing calculated results with experimental data, using the functionals BP and B3LYP, and the standard parametrization of a Coulomb-attenuated (range separated) version of B3LYP,⁴⁸ denoted here as CAM-B3LYP-A or short:CAM-A. In this parametrization, the fraction of Hartree–Fock exchange at large electronic separations approaches $\alpha + \beta = 0.65$.

The calculations of the uranium(VI) chloride fluorides $\text{UF}_{6-n}\text{Cl}_n$ employed optimized geometries reported by Straka and Kaupp,⁶⁰ which were obtained with the B3LYP functional and a scalar relativistic effective core potential (ECP). We used the uncontracted ANO-RCC basis set for uranium, with h functions removed in order to save computational resources. A justification for removing these functions in DFT calculations has been provided in ref 50. Since for the U(VI) complexes the main interest is in the ligand NMR shielding, a basis set devised for calculations of magnetic properties, namely the IGLO-III basis,⁹⁷ has been used for F and Cl. For comparison with previously published work on these systems, the calculations employed the BP and B3LYP functionals as well as a global hybrid with 50% HF exchange, BHLYP. The performance of range-separated functionals was further tested with a parametrization of CAM-B3LYP denoted here as CAM-B where $\alpha + \beta = 1$ (fully long-range corrected), and a fully long-range corrected hybrid form⁹⁸ of the Perdew–Burke–Ernzerhof (PBE) functional,⁹⁵ which is denoted here as LC-PBE0.

The computations involving the NWChem ZORA-SR g -shifts module were carried out for a set of small radicals for which geometries were optimized with ADF using ZORA/BP/TZVP with the exception of CdF, HgCN, and HgAg, where we used optimized bond lengths from ref 99. For the members of the test set with less heavy atoms, viz., CH_2 , CH_3 , CHO, HSiO, HSiS, SiOH, and SiSH, the IGLO-III basis set was used. For the molecules with heavier atoms, viz., TiF₃, CdF, HgF, HgCN, HgF, HgAg, and NpF₆, the uncontracted ANO-RCC basis was used for Ti, Cd, Hg, Ag, and Np (with h functions removed for Ti and Np and g functions removed for Cd, Hg, and Ag) and IGLO-III for the other atoms. A second set of g -shifts computations was performed using the 6-311G** basis set in place of IGLO-III. For comparison, we computed g shifts using ADF with the TZ2P basis set and the following methodologies: (i) a recently implemented second derivative g shift code for STO basis sets discussed in ref 35 and (ii) the van Lenthe, Wormer, and van der Avoird (LWA) implementation in ADF³⁴ which calculates the g tensor as a first derivative, with SO coupling included variationally in the ground state. For the g shift benchmark, the following functionals were employed: PBE, a global hybrid based on PBE with 25% HF exchange (PBE0),^{100,101} and CAM-A.

Default “extra fine” integration grids were used except for tellurium in tellurium complexes; uranium, fluorine, and chlorine in uranium halides; and neptunium and fluorine in NpF₆ where radial grids with additional points (350 radial grids in total) were used for better convergence of the numerical perturbation operator matrix elements. In the cases of HX, H₂X, and tellurium complexes, we checked the performance of the adopted numerical grid by comparing nonrelativistic shieldings obtained via analytical and numerical integration, respectively.

4. RESULTS AND DISCUSSION

4.1. Shielding of X in HX and H₂X. This section is mainly concerned with a validation of the ZORA-SR shielding implementation. Table 1 provides the relativistic effects obtained from scalar ZORA B3LYP computations of the heavy atom shielding constants for HX where X = F, Cl, Br, I, and At and for H₂X where X = O, S, Se, Te, and Po. The columns labeled as *nonrel* provide the calculated nonrelativistic shielding, which has been used to obtain the relativistic effects labeled as $\Delta(\text{rel})$ by taking the difference of the relativistic and the nonrelativistic shielding.

Table 1. Comparison of Scalar Relativistic Effects $\Delta(\text{rel})$ on Nuclear Shielding Obtained with NWChem^a (Implementation Described in Section 2) and ADF^b. B3LYP functional^c.

system ^d	σ_{iso} NWChem		σ_{iso} ADF	
	nonrel	$\Delta(\text{rel})^e$	nonrel	$\Delta(\text{rel})^e$
H ₂ O	326.5	−0.6	327.0	−0.6
H ₂ S	700.5	−4.9	700.4	−4.5
H ₂ Se	2084	−50.2	2084	−48.8
H ₂ Te	3547	−177.9	3549	−182.7
H ₂ Po	7080	−828.3	7080	−832.7
HF	411.3	−0.8	412.6	−0.8
HCl	938.1	−5.1	937.7	−5.0
HBr	2571	−44.0	2572	−43.2
HI	4433	−146.6	4434	−149.0
HAt	8451	−709.7	8453	−717.9

^a Scaled ZORA-SR with uncontracted ANO basis for X and TZVPP for H. ^b Scaled ZORA-SR with QZ4P STO basis set for X and TZ2P for H. NMR implementation of Wolff et al.²⁶ ^c Computed shielding of X for H₂X (X = O, S, Se, Te, Po) and HX (X = F, Cl, Br, I, At).

^d Equilibrium geometries for HX and H₂X taken from refs 77 and 78.

^e $\Delta(\text{rel}) = \sigma_{\text{iso}}^{\text{SR}} - \sigma_{\text{iso}}^{\text{nonrel}}$.

Although the ANO basis used for the heavy atoms X is not optimal for nonrelativistic calculations (in particular for the heaviest members of the series), its use in a fully uncontracted form should provide sufficient flexibility in order to obtain a reasonable estimate for the relativistic corrections. The same argument applies to the QZ4P Slater-type basis set used for the computations with ADF. Table S1 in the Supporting Information (SI) is similar to Table 1 but collects results obtained with the nonhybrid BP functional instead.

Overall, there is close agreement between the shielding constants calculated with the ZORA implementations in NWChem and those in the ADF program. The relativistic effects on the shielding constants generated by ZORA-SR are particularly close for the nonhybrid BP functional (see SI) but also very similar in the hybrid DFT calculations with the B3LYP functional. Differences between the two codes must be expected, most importantly due to the different nature of the basis sets (GTO versus STO, and details on how these basis sets were optimized) and the numerical integration grids used to calculate DFT-related XC contributions. Further, the HF exchange contributions in ADF are calculated via an orbital-pair density fitting procedure and therefore affected by the quality of the auxiliary density-fit basis sets that accompany the STO basis sets included in the ADF basis set library. Another contributing factor is how the potential in the denominator in \mathcal{K} , eq 5, is approximated: NWChem uses a sum of atomic nuclear and Hartree potentials, whereas ADF further includes a sum of approximate exchange–correlation potentials. Keeping these technical differences in mind, the agreement of the results obtained with the two ZORA implementations can be considered as excellent, showing that comparable relativistic effects in the shielding constants are generated. Considering, further, that calculations of chemical shifts tend to furnish cancellation of systematic errors such as those from numerical integration, the new implementation in NWChem is expected to predict very comparable chemical shifts to the one in ADF when using the same functionals and basis sets of comparable flexibility.

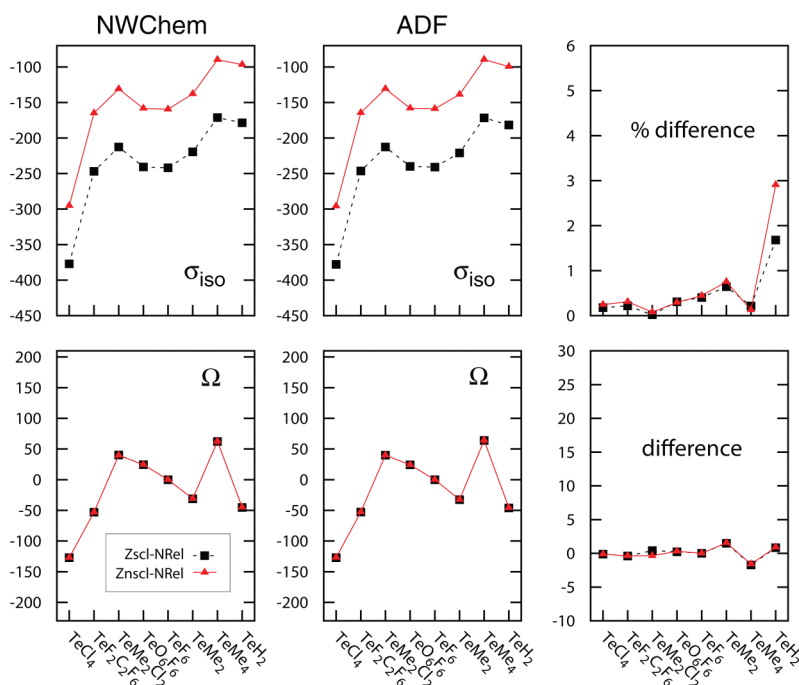


Figure 1. Comparison of relativistic effects (rel – nonrel) on ^{125}Te shielding constants and shielding tensor spans (Ω), in ppm. Calculations using the ZORA-SR shielding implementations in NWChem reported in this work, and the implementation by Wolff et al.²⁶ in the ADF program. Unscaled and scaled ZORA-SR, B3LYP functional.

Manninen et al.¹⁰² have previously calculated relativistic effects of order c^{-2} , corresponding to the Breit–Pauli operator, on nuclear shielding of the HX and H_2X series directly as perturbations, using nonrelativistic HF and complete active space (CAS) MCSCF wave functions. The effects were divided into passive and active contributions, with the former originating from relativistic changes of the wave function and the latter from relativistic terms $\mathcal{O}(c^{-2})$ in the magnetic perturbation operators. A comparison of our results with those of Manninen et al. shows the following: (i) The nonrelativistic shieldings differ by about 3–7%, which must be attributed to basis set effects and the different treatment of correlation and exchange in the calculations. (ii) Regarding the relativistic effects on the X shielding, we considered the “passive” SR terms of Table 1 of ref 102 and a scalar relativistic (SR) “active” term from Table 3 but no terms involving the Fermi-contact and spin-dipole operators. The ZORA analogs of the latter contribute to the shielding in variational relativistic calculations with spin–orbit coupling included and will be the subject of a follow-up study. The sum of several passive scalar relativistic terms was calculated to be positive by Manninen et al. in ref 102 but shown to be sensitive to the basis set—in particular the augmentation with high exponents—and correlation. The sensitivity arises in part from a cancellation of perturbation terms of opposite sign, which in some cases individually exceed the total by more than an order of magnitude. An active orbital–Zeeman–kinetic energy term was calculated to be positive for the HX and H_2X series and shown to be relatively insensitive to correlation. The SR effects on the X shielding are calculated to be negative with ZORA DFT. However, Manninen et al. noted in ref 102 that a large negative active SR contribution to the shielding was neglected in the study. This and other previously neglected terms were later calculated explicitly, leading to additional large and negative SR contributions to the X shielding in the HX and H_2X series, which overpower the positive terms.¹⁰³

Given the sensitivity to correlation and the fact that the present ZORA calculations are variational, a direct comparison with the data by Manninen et al. is difficult. We therefore deem the excellent agreement with the ADF ZORA implementation and the overall sign of the SR effects on the X shielding as sufficient to validate the method. It is noted in passing that a set of perturbation terms corresponding to the Pauli approximation can be obtained from ZORA by using $\mathcal{H} \approx 1 + V/(2c^2)$ and applying partial integration or the turnover rule for the momentum operator in the matrix elements.

4.2. NMR Shielding and Chemical Shifts of Tellurium Compounds. Calculations for a set of selected tellurium compounds employed X-ray geometries taken from several references indicated in the Computational Details section and in the table footnotes. Figure 1 displays a graphical comparison of the scalar relativistic effects on isotropic Te shielding constants σ_{iso} and on the shielding tensor span Ω , calculated with the NWChem ZORA-SR implementation and with the ADF package using the B3LYP functional. The corresponding numerical data for the B3LYP and the BP functional, as well as a graphical comparison similar to that of Figure 1 but for BP computations, are provided in Tables S2 and S3 and in Figure S1 of the Supporting Information (SI). The calculations employed an unscaled as well as a scaled ZORA formalism, as outlined near the end of section 2.

The agreement between the NWChem and the ADF ZORA results seen in Figure 1 and the corresponding tables and the figure in the SI is very good. Further, the effects from the ZORA scaling are also very close. Reasons have already been provided in the previous section why minor differences must be expected between the calculations performed with NWChem and ADF. The results for the Te compounds show that the combination of the uncontracted ANO-RCC basis for the heavy atom with 6-311G* for lighter atoms provides a suitably flexible basis set

Table 2. ^{125}Te Chemical Shifts (ppm), ZORA-SR Calculations with NWChem^a and Experimental Data

system	BP		B3LYP		CAM-B3LYP		ref 83 ^d	experimental	
	opt ^b	X-ray ^c	opt ^b	X-ray ^c	opt ^b	X-ray ^c			
TeCl ₄		1703		1612		1835	1541	1725	ref 104
TeC ₂ F ₆	1379		1470		1804		1404	1368	ref 105
TeF ₂ C ₂ F ₆	959	864	867	792	1256	1220		1187	refs 106 and 107
TeMe ₂ Cl ₂	494	432	458	398	506	466	451	734	ref 108
TeO ₆ H ₆	570	598	498	526	928	1008	614	707	ref 109
TeF ₆	416	434	327	342	750	811	604	543	ref 110
TeEt ₂	410		399		558		346	380	ref 111
TeMe ₂	0	0	0	0	0	0	0	0	
TeMe ₄	−292	−185	−326	−212	−303	−136	−280	−67	ref 112
TeH ₂	−1008	−878	−964	−835	−983	−846	−770	−621	refs 111 and 113
Δ ^e	173	177	218	226	242	182	131		

^a Scaled ZORA-SR with uncontracted ANO basis set for Te and 6-311G* basis set for C, H, F, and Cl. ^b Scalar ZORA optimized geometries obtained with ADF using the BP functional and the TZ2P basis set. ^c X-ray geometries obtained from ref 86 for TeCl₄, ref 87 for TeF₂C₂F₆, ref 88 for TeMe₂Cl₂, ref 89 for TeO₆H₆, ref 90 for TeF₆, ref 91 for TeMe₂, ref 92 for TeMe₄, ref 93 for TeH₂. ^d Quasirelativistic generalized Moller–Plesset perturbation theory (QR-GUMP2). Electron correlation included via MP2. The X-ray geometries for TeCl₄, TeMe₂Cl₂, TeO₆H₆, and TeH₂ are the same geometries used in our calculations. ^e Δ = unsigned mean deviation from experimental results.

for this type of calculation that should allow for routine applications in systems with one or a few truly heavy atoms (in the context of this and the following sections, atoms with nuclear charges less than about 20 to 30 may be considered light atoms). For calculations of ligand NMR shifts, it would be advisable to use basis sets designed for NMR shielding calculations for the ligand atoms. The comparison of unscaled and scaled ZORA results indicates that the main effect of the ZORA scaling is a relatively constant positive change of the isotropic shielding, which cancels to a large degree when evaluating chemical shifts. The effects on the tensor span are hardly noticeable on the scale of the plots. Noticeable effects on chemical shifts due to ZORA scaling are likely to be significant only for systems where the scaling factors deviate noticeably from unity for valence orbitals that are contributing to the shielding.

Table 2 collects tellurium chemical shifts calculated with optimized geometries and with geometries determined experimentally from X-ray diffraction. The tellurium chemical shifts, δ_{Te} , were obtained with dimethyl telluride, TeMe₂, as the reference, using $\delta_{\text{Te}} = \sigma_{\text{TeMe}_2} - \sigma_{\text{compound}}$. Table S4 of the SI provides the shielding constants calculated for the reference. Calculations were performed with the functionals BP, B3LYP, and CAM-B3LYP. The last column in the table lists corresponding experimental chemical shifts obtained in varying solvents. Solvent effects together with thermal motions have been estimated to contribute on the order of ± 100 ppm to the experimental chemical shifts.^{85,114} The data reveal several trends: (i) For the compounds with positive chemical shifts, there is a tendency for $\delta_{\text{Te, B3LYP}} < \delta_{\text{Te, BP}} < \delta_{\text{Te, exp}} < \delta_{\text{Te, CAM}}$. For example, for TeO₆H₆ for the optimized geometry, $\delta_{\text{Te, B3LYP}} = 498$ ppm, $\delta_{\text{Te, BP}} = 570$ ppm, $\delta_{\text{Te, exp}} = 707$ ppm, and $\delta_{\text{Te, CAM}} = 928$ ppm. It appears that CAM-B3LYP is overestimating these chemical shifts in the framework of scalar ZORA. (ii) For the compounds with negative chemical shifts, the X-ray geometries give results closer to experiment than the optimized geometries. For example, for TeMe₄, the closest agreement with experiment is $\delta_{\text{Te, CAM}} = -136$ ppm as opposed to the “best” result based on the optimized geometry, $\delta_{\text{Te, BP}} = -292$ ppm. (iii) The deviations between the DFT results and the quasi-relativistic MP2 data of ref 83 (which

includes spin–orbit effects) are roughly of the same magnitude as deviations between the various sets of calculations and experimental results.

A comparison of calculated and experimental chemical shifts is provided in Figure 2. Overall, the predicted chemical shifts correlate reasonably well with experimental data for all three functionals. For most of the tellurium compounds, the predicted scalar relativistic CAM-B3LYP chemical shifts are slightly further away from experimental values than the corresponding nonrelativistic chemical shifts. Mean unsigned deviations of the scalar ZORA results from experimental results are listed in Table 2. The deviations range from 7% (BP functional) to 10% (CAM-B3LYP with optimized geometries) of the chemical shift range of the set of Te compounds, while the deviations for the MP2 literature data are slightly smaller on average (6% of the chemical shift range). The CAM-B3LYP data calculated with the experimental geometries have similar deviations from experimental results as both sets of BP functional data. On the scale of the chemical shift range, the overall performance of the different functionals is roughly comparable. Substantial and systematic improvements of the DFT calculations should further include solvent effects and spin–orbit terms in the shielding tensors.

4.3. Shielding of Uranium(IV) Chloride Fluoride Complexes. Table 3 collects scalar relativistic ¹⁹F shielding constants for a set of uranium(VI) halide complexes. The fluorine shieldings calculated with the BP and B3LYP functionals are very close to those obtained by Straka and Kaupp,⁶⁰ which are listed in Table 3 in parentheses. Straka and Kaupp employed a small-core relativistic effective core potential (RSC ECP) as opposed to our all-electron scalar ZORA approach. Since the calculations are performed for the shielding of a light nucleus in heavy atom compounds, the two approaches are expected to yield comparable results. It is reassuring that this is indeed the case. We note that the present ZORA-SC computations were carried out with B3LYP/RSC-TZVP optimized geometries taken from ref 60. The optimized bond lengths are very similar to those reported by Schreckenbach.⁶² Table 3 also lists results obtained by Straka and Kaupp with the B3LYP functionals, which overall agreed best

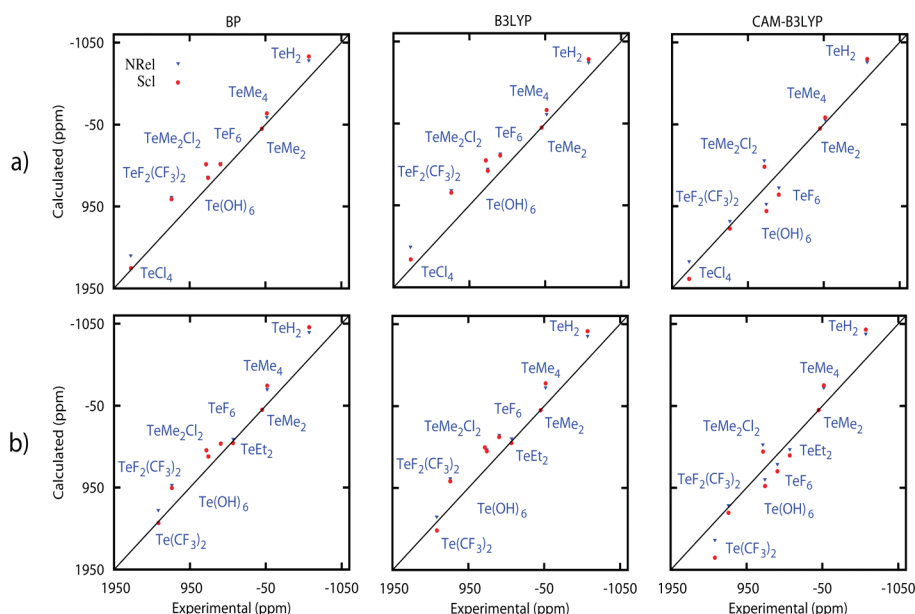


Figure 2. Performance of BP, B3LYP, and CAM-B3LYP functionals for ^{125}Te chemical shifts calculated (a) with X-ray crystallographic geometries and (b) with optimized geometries. NWChem, scaled ZORA-SR. The 45° lines indicate where $\delta_{\text{calcd}} = \delta_{\text{expt}}$.

Table 3. Dependence of ^{19}F Shielding in $\text{UF}_n\text{Cl}_{6-n}$ ($n = 1-6$)^a on the XC Functional^b

system	position ^c	BP	B3LYP	BHLYP	BHH	CAM-A	CAM-B	LC-PBE0	experimental ⁶⁰		
UF_6	F_{TF}	-720.3	(-727.0)	-697.9	(-706.2)	(-606.7)	-604.4	-644.6	-604.1	-611.8	-575.3
UF_5Cl	F_{TF}	-712.0	(-714.6)	-689.2	(-692.6)	(-606.2)	-612.8	-647.1	-614.2	-628.2	-573.3
<i>trans</i> - UF_4Cl_2	F_{TF}	-706.8	(-705.9)	-685.0	(-684.2)	(-611.1)	-625.2	-652.0	-625.7	-645.5	-566.8
<i>cis</i> - UF_4Cl_2	F_{TF}	-705.0	(-705.1)	-682.5	(-682.7)	(-606.7)	-619.4	-648.4	-621.2	-640.2	-571.3
<i>mer</i> - UF_3Cl_3	F_{TF}	-701.7	(-699.7)	-680.3	(-678.0)	(-613.0)	-630.9	-653.1	-631.0	-654.7	-564.3
<i>trans</i> - UF_2Cl_4	F_{TF}	-699.0	(-695.9)	-679.3	(-675.6)	(-620.0)	-641.2	-657.1	-639.1	-666.5	-557.4
UFCl_5	F_{tCl}	-652.5	(-647.8)	-659.9	(-655.5)	(-645.2)	-664.6	-652.0	-644.7	-675.6	-585.6
UF_3Cl	F_{tCl}	-679.6	(-682.0)	-683.2	(-689.0)	(-629.0)	-625.6	-644.7	-612.8	-623.8	-592.8
<i>cis</i> - UF_4Cl_2	F_{tCl}	-668.0	(-667.6)	-671.2	(-672.9)	(-627.5)	-632.3	-643.3	-619.5	-636.7	-597.1
<i>mer</i> - UF_3Cl_3	F_{tCl}	-662.0	(-659.1)	-666.5	(-664.7)	(-635.8)	-647.7	-647.7	-631.3	-654.4	-593.9
<i>fac</i> - UF_3Cl_3	F_{tCl}	-659.9	(-657.6)	-663.4	(-662.6)	(-628.2)	-638.8	-643.2	-625.5	-647.6	-597.7
<i>cis</i> - UF_2Cl_4	F_{tCl}	-655.6	(-651.8)	-660.9	(-657.8)	(-636.7)	-652.6	-647.7	-635.8	-662.8	-592.3

^a Using scaled scalar ZORA with uncontracted ANO basis set for uranium and IGLO-III for fluorine and chlorine. ^b The data in parentheses correspond to calculations reported in ref 60. CAM-A = CAM-B3LYP-A, original parametrization ($\alpha + \beta = 0.65$), CAM-B = CAM-B3LYP-B, fully long-range-corrected parametrization ($\alpha + \beta = 1$), BHH = BeckeHandH, 50% HF exchange + 50% Slater exchange + 50% PW91LDA correlation. ^c F_{TF} = fluorine *trans* to another fluorine. F_{tCl} = fluorine *trans* to chlorine.

with experimental results. Because of the good agreement of our BP and B3LYP data with those of Straka and Kaupp, we decided to forego additional BHLYP computations and focus on the range-separated functional instead. Results from calculations with another 50% global hybrid, the Becke “half and half” parametrization (BHH), are additionally listed in Table 3. A visual comparison of the results with experimental data is shown in Figure 3. Experimental absolute shieldings were taken from ref 60. The graph in Figure 3 has been designed to appear similar to a graph in ref 60, which we found very suitable for illustrating the various trends in the data set.

As outlined in the Computational Details, the performance of three long-range corrected functionals has been tested. Among those, CAM-B3LYP in its fully long-range corrected form (CAM-B) gives calculated shieldings that are closest to experimental results, with LC-PBE0 being close in some cases and

more similar to CAM-A in others. For the example mentioned in the Introduction, *mer*- UF_3Cl_3 , the CAM as well as the LC-PBE0 calculations yield F_{tCl} and F_{TF} to be approximately equal. This is an improvement over BP and B3LYP, which give substantially smaller magnitudes for F_{tCl} vs F_{TF} , whereas the experimental data show the opposite trend. The improvement with the long-range corrected functionals is not as strong as it is with BHLYP, which, for *mer*- UF_3Cl_3 , gives a 23 ppm more positive shielding for F_{TF} vs F_{tCl} . Our calculations with the Becke half-and-half functional also reproduce this trend. Overall, the BP and B3LYP functionals appear to somewhat better reproduce the trend visible in Figure 3 within the F_{TF} group, unlike BHLYP and the range-separated functionals. However, BP and B3LYP fail to produce the overall larger shielding magnitudes in this group compared to F_{tCl} .

Beyond a comparison of their performance, a detailed analysis of how exactly the various approximations in the XC functionals

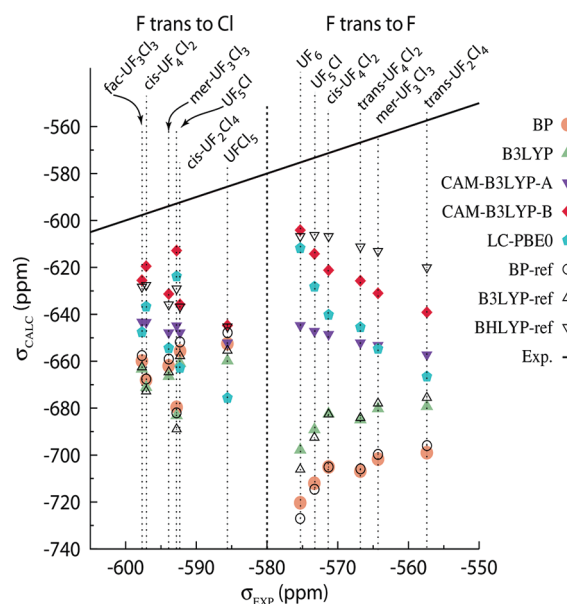


Figure 3. Performance of different functionals in the calculation of ^{19}F shielding in the $\text{UF}_n\text{Cl}_{6-n}$ ($n = 1-6$) series. BP-ref, B3LYP-ref, and BHLYP-ref are results from ref 60. The calculations used scaled ZORA-SR.

applied in this work influence the ^{19}F shieldings in the U(VI) halides is beyond the scope of this work. The comparison of the functionals already indicates that the fraction of HF exchange in the functional, and the balance of electron correlation with self-interaction and asymptotic behavior of the potential that it entails, influences various trends within the data set and the overall magnitude of the absolute shielding in different ways. A full long-range correction is evidently not beneficial in conjunction with the XC functionals applied in this study. B3LYP appears to have a too small fraction of HF exchange to provide a substantial difference over the nonhybrid BP functional for the U(VI) halides. Given the fact that the global 50% hybrid BHLYP previously gave the best overall performance, giving larger magnitudes of F_{TF} versus F_{TCI} shieldings, and considering that the standard parametrization of CAM-B3LYP only goes to 65% HF exchange at large interelectronic distances, one might have expected the CAM-A functional to perform well for this challenging set of chemical shifts. However, while the shielding constants overall are closer to experimental results (compared to BP and B3LYP), a clear trend for F_{TCI} vs F_{TF} is not obtained with this parametrization. Spin-orbit effects were previously shown to be too small to reconcile the differences between nonhybrid DFT calculations and experimental results.⁶² Spin-orbit and other relatively minor effects may help to close the remaining gaps between some of the hybrid DFT calculations and experimental results. An evaluation of range-separated functionals in spin-orbit ZORA NMR calculations, and the inclusion of solvent effects, is planned for a follow-up study.

Schreckenbach has previously provided theoretical estimates for ^{235}U chemical shifts⁶² based on all-electron scalar ZORA computations with nonhybrid functionals as implemented in ADF, indicating a large shift range that may exceed 20 000 ppm. For completeness, Table S5 in the Supporting Information provides the uranium shielding constants and tensor spans for the $\text{UF}_n\text{Cl}_{6-n}$ series. The agreement of the nonhybrid BP results

with those from Schreckenbach's paper is quite close, which may serve as further validation of the NWChem implementation. The shielding constants are clearly dominated by negative paramagnetic contributions. The results serve as benchmark data only since physically meaningful absolute shieldings for U should be calculated with spin-orbit coupling and with a finite nucleus model. Regarding different functionals, we note that with the range-separated hybrid functionals, and in particular with the full LR corrected versions, the shielding constants become significantly more positive.

4.4. g -Shifts of Small Radicals. Data in Table 4 illustrate the performance of the NWChem ZORA g shift module for predicting g -shifts for a set of selected small radicals. A comparison is made with a second-order ZORA implementation in ADF that was recently reported in ref 35 ("AP", spin-orbit coupling treated as a perturbation, similar to the present NWChem implementation). Further, a comparison is made with ADF ZORA calculations using a first-order derivative approach developed by van Lenthe, Wormer, and van der Avoird (LWA). In these calculations, spin-orbit coupling is included variationally in the ground state calculations, see also ref 64. The TZ2P basis set has been used for the ADF calculations since acceptable agreement with experimental results has been obtained in previous works using this STO basis.^{34,35,99} The test set allows for a comparison of the influence of nuclear charges (lighter to heavier elements) and different bonding situations. Table 4 compares different types of functionals and includes results obtained with the standard parametrization of CAM-B3LYP (CAM-A). The agreement of the second-order and first-order ZORA approaches is reasonably close except for the heavy atomic systems with rotational symmetry. This behavior is expected and has been discussed in detail elsewhere.^{55,65,99} The second-order approaches tend to underestimate the magnitude of the isotropic g shift for such systems because the tensor component parallel to the rotational axis is suppressed. TiF_3 is one of the systems where hybrid functionals offer a clear improvement, in line with previous studies.¹⁰ Atoms with tightly bound lone pairs (fluorine) and the notoriously difficult treatment of response properties in systems with 3d metals tend to benefit from hybrid functionals. For NpF_6 , the g shift also increases significantly in magnitude when going from the nonhybrid GGA to the hybrid functionals. In the NWChem calculations, this brings the result closer to the experimental g shift derived from solid state data.¹²² Regarding the performance of CAM-B3LYP, in comparison with the changes when going from PBE to PBE0 and in comparison with the differences between calculations and experimental results, the results are in most cases very similar to those obtained with the global hybrid PBE0. Overall, the range-separated hybrid does not appear to outperform the PBE0 global hybrid for the g -shifts of the few-atomic radical test set of Table 4. During the course of this study, we noticed that the diamagnetic g -shifts exhibited some sensitivity to the approximations used for \mathcal{H} (eq 5). For example, for the molecule HSiS , the diagonal elements of the diamagnetic g shift tensor calculated with NWChem were $(-0.0154, -0.0347, -0.0147)$ ppt (BP functional, unscaled ZORA), whereas the calculation with ADF gave $(-0.0359, -0.0626, -0.0400)$ ppt. It was found that the difference is most likely due to the exclusion of the atomic XC potentials in V in the construction of \mathcal{H} in NWChem, unlike in ADF where they are included. A comparison was made between the two codes, using in both cases only the nuclear potentials in the construction of \mathcal{H} . In this case, the diagonal elements of the diamagnetic g shift tensor for molecule HSiS were calculated as $(-0.475, -0.552, -0.528)$ ppt and

Table 4. Calculated Δg Shifts for Selected Molecules in ppt^a

system ^c	NWChem			ADF (AP) ^b		ADF (LWA) ^b	exptl.
	PBE	PBE0	CAM-A	PBE	PBE0	PBE	
CH ₂	0.1824	0.1674	0.1819	0.2057	0.1917	0.1383	
CH ₃	0.5084	0.4763	0.5160	0.5380	0.5080	0.4733	0.10 ^d
CHO	-3.028	-3.084	-3.184	-2.333	-2.298	-2.442	-2.1 ^e
HSiO	-1.704	-1.552	-1.440	-1.696	-1.539	-1.820	-1.3 ^f
HSiS	-1.526	-1.322	-0.2505	-1.646	-1.463	-1.895	
SiOH	-23.83	-25.18	-35.47	-24.13	-25.83	-24.13	
SiSH	-13.57	-14.44	-17.93	-14.03	-15.12	-14.09	
TiF ₃	-29.20	-41.26	-43.44	-28.16	-39.94	-28.54	-77.92 ^g
CdF	-11.92	-10.84	-10.69	-11.50	-10.45	-10.32	-9.667 ^h
HgH	-97.85	-95.05	-99.82	-93.84	-91.66	-131.0	-125 ⁱ
HgCN	-55.25	-55.87	-57.97	-53.76	-54.58	-78.13	
HgF	-17.21	-12.86	-10.09	-17.48	-13.02	-31.55	-30.7 ^j
HgAg	-64.96	-64.18	-61.20	-61.58	-60.56	-80.28	-59.00 ^k
NpF ₆	-1795	-2359	-2299	-2761	-3015	-2564	-2606 ^l

^a Result in parts per thousand. NWChem calculations with uncontracted ANO basis for Ti, Cd, Hg, Ag, and Np and IGLO-III for other atoms. ADF calculations with STO basis set TZ2P. Both calculations used unscaled ZORA-SR. CAM-A = CAM-B3LYP, original parametrization ($\alpha + \beta = 0.65$). AP = ADF calculations with ZORA second-derivative approach similar to the present NWChem implementation, developed by Autschbach and Pritchard.³⁵ LWA = ADF calculations with ZORA first-order approach including spin-orbit coupling variationally, as developed by van Lenthe et al.³⁴ ^b CdF, HgCN, and HgAg were calculated using AP and LWA implementations in ADF. Data for the other molecules were taken from ref 35. ^c All systems calculated as doublets except triplet CH₂. Optimized geometries obtained with ADF using ZORA/BP/TZ2P with the exception of CdF, HgCN, and HgAg, for which optimized bond lengths were taken from ref 99. ^d Ref 115. ^e Ref 116. ^f Ref 117. ^g Ref 118. ^h Ref 119. ⁱ Ref 120. ^j Ref 121. ^k Ref 122. ^l Ref 123. Regarding NpF₆, see text and SI for further details.

(-0.469, -0.556, -0.521) ppt with NWChem and ADF, respectively. These results agree much better. Due to the very small magnitude of Δg^{dia} , the overall results are not strongly impacted by this sensitivity.

The comparison of the NWChem GTO basis calculations to the ADF results calculated with an STO basis set shows very similar trends for the functionals and overall a close agreement, with the exception of NpF₆. Calculations performed for this system with only the nuclear potential used in the construction of \mathcal{K} produced significantly more negative g -shifts (changes on the order of -10^3 ppt, PBE functional) but the differences between the GTO and STO basis calculations remained. Table S6 of the Supporting Information is similar to Table 4 except that the NWChem results were obtained with the 6-311G** basis in place of IGLO-III for light atoms. The comparison of the data in the two tables shows that the GTO basis results are very close. Additional calculations with the PBE0 functional and the GTO basis sets 6-311++G** and aug-cc-pVDZ for fluorine yielded an isotropic g shift of -2370 and -2357 ppt, respectively, for NpF₆. These values are very close to the g -shifts listed in Tables 4 and S6. Hence, these calculations do not indicate a very pronounced fluorine basis set dependence of the NpF₆ g shift. Further, a calculation for NpF₆ (PBE functional) gave differences of only a few parts per thousand upon removal of the g functions from the Np ANO basis. Calculations with varying fluorine STO basis sets, going from an unpolarized double- ζ (DZ) to a large triply polarized quadruple- ζ basis (QZ4P), caused changes of less than -100 ppt. Overall, calculations with varying quality STO and GTO basis sets of varying sizes did not produce large enough changes to reconcile the g shift of NpF₆ calculated with the different codes.

We note that STOs used reach further into the asymptotic region than GTOs, and therefore the coverage of more diffuse regions by the basis set might be of some significance. The

addition of diffuse p functions to the IGLO-III fluorine GTO basis did, however, not significantly change the calculated NpF₆ g factor. The TZ2P basis used for the STO calculations has one set of d and f polarization functions each for fluorine. In comparison, the IGLO-III basis for fluorine has two sets of d functions but no f . A test calculation with the TZVPP GTO basis for fluorine, which includes an f polarization function, gave a g shift of -1806 ppt (PBE functional). This result is not much different from the IGLO-III value of Table 4. A characterization of the frontier orbitals showed that the NWChem and ADF calculations converged to the same ground state.

The GTO based NWChem implementation does not approximate the Coulomb and exchange Fock matrix contributions beyond the approximations posed by the finite GTO basis set (and the density functional). On the other hand, in the ADF calculations, the Coulomb and XC potential and the HF exchange matrix are calculated by using an auxiliary density fitting basis which is limited to angular momenta of $l \leq 4$. Since the spin density of NpF₆ stems from the formal Np $5f^1$ configuration, the limitation of the angular momentum is of more concern for this system than for the other molecules in the test set. For the Coulomb and HF exchange, there is presently no alternative way of performing the ADF calculations. For the pure-DFT XC Fock matrix elements, however, it is possible to use the "exact" density represented in the AO basis instead of the fitted density. A series of such calculations (Table S7, SI) showed a significant decrease in the magnitude of the NpF₆ g factor, with the STO basis results obtained this way being closer to the GTO basis results. Quantitative agreement may therefore be expected if the fit-related limitations for the Coulomb and exact exchange Fock matrices are lifted. At present, the hybrid DFT GTO data obtained with the implementation reported herein (section 2) must be considered as the most reliable ZORA results.

5. SUMMARY AND OUTLOOK

Comparisons between ZORA calculations performed with the ADF code and Slater-type orbitals, and the new Gaussian basis ZORA NMR/EPR implementation in NWChem, using nonhybrid and standard global hybrid functionals, have shown good agreement and therefore serve as a validation of the new implementation. The calculation of the required ZORA perturbation operator matrix elements by numerical integration is rather straightforward. For *g*-shifts, good performance was found both for the PBE0 hybrid (25% HF exchange globally) and for the CAM-B3LYP range-separated hybrid. Range-separated hybrid functionals show some promise for NMR shielding calculations for difficult situations such as the ¹⁹F shielding in the UF_{6-n}Cl_n series of complexes, but it is possible that a universal setting for the range-separation parameter is limiting the performance of this class of functionals. The obvious next step is to calculate the fluorine shieldings of these systems with range-separated hybrids within a spin-orbit ZORA NMR framework, and with inclusion of solvent effects in order to more closely model the experimental conditions. Developments to enable spin-orbit ZORA NMR computations are currently under way in our team. Although previous work on the U(VI) halide series did not indicate that spin-orbit effects are able to bridge the large gap between nonhybrid DFT results and experimental results,⁶² they may further improve results with range-separated functionals and global hybrids with high fractions of HF exchange. As pointed out already, the agreement with experimental results is not ideal.

APPENDIX: ATTENUATED TWO-ELECTRON INTEGRALS USING THE RYS SCHEME

For completeness, we provide the details for the evaluation of the two-electron integrals in conjunction with range-separated functionals. In the range-separated approach, the electron repulsion is separated into long- and short-range parts. Since the NMR implementation⁴⁰ in NWChem is based on the Rys quadrature method,^{123,124} we present the necessary changes for the modified two-electron integrals for attenuated interactions within this approach. The general form of the four-center, two-electron electron repulsion integral (ERI) is defined as

$$\left(\chi_i\chi_j\left|\frac{1}{r_{12}}\right|\chi_k\chi_l\right) = \int \int \chi_i(\mathbf{r}_1)\chi_j(\mathbf{r}_1)\frac{1}{r_{12}}\chi_k(\mathbf{r}_2)\chi_l(\mathbf{r}_2)d^3r_1d^3r_2 \quad (\text{A-1})$$

where the $\chi_{i,j,k,l}$ are Gaussian primitive basis functions. Within the Rys scheme, the above integral is redefined as follows:

$$\left(\chi_i\chi_j\left|\frac{1}{r_{12}}\right|\chi_k\chi_l\right) = \frac{2}{\sqrt{\pi}}\int_0^\infty du(\chi_i\chi_j|e^{-u^2r_{12}}|\chi_k\chi_l) \quad (\text{A-2})$$

and similarly, the attenuated ERI can be written as

$$\left(\chi_i\chi_j\left|\frac{\text{erf}(\gamma r_{12})}{r_{12}}\right|\chi_k\chi_l\right) = \frac{2}{\sqrt{\pi}}\int_0^\gamma du(\chi_i\chi_j|e^{-u^2r_{12}}|\chi_k\chi_l) \quad (\text{A-3})$$

by using $\text{erf}(\gamma r_{12}) = \int_0^{\gamma r_{12}} e^{-t^2} dt$ and a variable change, $u = t/\gamma$. After a further change of variables $t^2 = u^2/(\rho + u^2)$ and

$1/\rho = 1/(a_i + a_j) + 1/(a_k + a_l)$, eq A-2 becomes

$$\left(\chi_i\chi_j\left|\frac{1}{r_{12}}\right|\chi_k\chi_l\right) = \int_0^1 P_L(t) e^{-Xt^2} dt \quad (\text{A-4})$$

where P_L is a polynomial of degree L in t^2 and $X = \rho|r_A - r_B|^2$, and A and B are the centers of the orbital products ij and kl , respectively. X implicitly depends on the Gaussian exponents a_i, a_j, a_k , and a_l via ρ and the coordinates of the centers of the four Gaussians involved in the integral. The integral over t can be evaluated using an N point quadrature formula such as

$$\left(\chi_i\chi_j\left|\frac{1}{r_{12}}\right|\chi_k\chi_l\right) = \sum_{\alpha=1}^N P_L(t_\alpha)W_\alpha \quad (\text{A-5})$$

where W_α is a weight factor. Similarly, the attenuated ERI in this approach becomes

$$\left(\chi_i\chi_j\left|\frac{\text{erf}(\gamma r_{12})}{r_{12}}\right|\chi_k\chi_l\right) = \int_0^\beta P_L(t) e^{-Xt^2} dt \quad (\text{A-6})$$

where $\beta = \gamma/(\rho + \gamma^2)^{1/2}$. In order to use the Rys formulas directly, we need another transform $t' = t/\beta$, to obtain

$$\left(\chi_i\chi_j\left|\frac{\text{erf}(\gamma r_{12})}{r_{12}}\right|\chi_k\chi_l\right) = \int_0^1 P_L(t'\beta) e^{-X\beta^2 t'^2} \beta dt' \quad (\text{A-7})$$

which can be mapped directly onto eq A-5 with suitably modified weights and variables.

ASSOCIATED CONTENT

S Supporting Information. X shielding constants for the HX series and tellurium shielding tensor data calculated with the BP functional. Numerical data used for Figure 1. Analog of Figure 1 but for the BP functional. Te reference shielding constants. Calculated U shielding tensor data. *g*-shifts calculated with different basis sets. This information is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jochena@buffalo.edu.

ACKNOWLEDGMENT

This work has received financial support from the U.S. Department of Energy, grant no. DE-SC0001136 (BES Heavy Element Chemistry Program). We thank the Center for Computational Research (CCR) at the University at Buffalo for continuing support of our research projects. A portion of the research was performed using EMSL, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research located at Pacific Northwest National Laboratory (PNNL). PNNL is operated for the Department of Energy by the Battelle Memorial Institute under Contract DE-AC06-76RLO-1830. N.G. thanks M. Dupuis and W. A. de Jong for useful discussions and would like to acknowledge the DOE BES Heavy Element Chemistry Program (PI: De Jong, PNNL) of the

U.S. Department of Energy, Office of Science, and NWChem development for support.

REFERENCES

- (1) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293–352.
- (2) Malkin, V. G.; Malkina, O. L.; Eriksson, L. A.; Salahub, D. R. The calculation of NMR and ESR spectroscopy parameters using density functional theory. In *Modern Density Functional Theory: A Tool for Chemistry*; Seminario, J. M., Politzer, P., Eds.; Elsevier: Amsterdam, 1995; Vol. 2.
- (3) Facelli, J. C. Shielding Calculations. In *Encyclopedia of Nuclear Magnetic Resonance*; Grant, D. M., Harris, R. K., Eds.; John Wiley & Sons: Chichester, U. K., 2002; Vol. 9.
- (4) Kaupp, M. Ab initio and density functional calculations of electronic g-tensors for organic radicals. In *EPR Spectroscopy of Free Radicals in Solids. Trends in Methods and Applications*; Lund, A., Shiotani, M., Eds.; Kluwer: Dordrecht, The Netherlands, 2002.
- (5) Autschbach, J.; Ziegler, T. *Coord. Chem. Rev.* **2003**, *238/239*, 83–126.
- (6) Autschbach, J. Relativistic effects on magnetic resonance parameters and other properties of inorganic molecules and metal complexes. In *Relativistic Methods for Chemists*; Ishikawa, J., Barysz, M., Eds.; Springer: London, 2010; Vol. 10.
- (7) Kaupp, M.; Bühl, M.; Malkin, V. G. *Calculation of NMR and EPR Parameters. Theory and Applications*; Wiley-VCH: Weinheim, Germany, 2004.
- (8) Rieger, P. H. *Electron spin resonance. Analysis and interpretation*; The Royal Society of Chemistry: Cambridge, U.K., 2007; p 3.
- (9) Moon, S.; Patchkovskii, S. First-principles calculations of paramagnetic NMR shifts. In *Calculation of NMR and EPR Parameters. Theory and Applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004.
- (10) Kaupp, M.; Köhler, F. H. *Coord. Chem. Rev.* **2009**, *253*, 2376–2386.
- (11) Bertini, I.; Turano, P.; Vila, A. J. *Chem. Rev.* **1993**, *93*, 2833–2932.
- (12) Rastrelli, F.; Bagno, A. *Chem.—Eur. J.* **2009**, *15*, 7990–8004.
- (13) Autschbach, J.; Zheng, S. *Annu. Rep. NMR Spectrosc.* **2009**, *67*, 1–95.
- (14) Nomura, Y.; Takeuchi, Y.; Nakagawa, N. *Tetrahedron Lett.* **1969**, *8*, 639–642.
- (15) Kaupp, M.; Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *J. Am. Chem. Soc.* **1995**, *117*, 1851–1852; Erratum *ibid.*, p. 8492.
- (16) Schreckenbach, G.; Wolff, S. K.; Ziegler, T. Covering the entire periodic table: Relativistic density functional calculations of NMR chemical shifts in diamagnetic actinide compounds. In *Modeling NMR Chemical Shifts*; Facelli, J. C., de Dios, A. C., Eds.; American Chemical Society: Washington, DC, 1999; ACS Symposium Series 732.
- (17) Kaupp, M.; Malkin, V. G.; Malkina, O. L. NMR of transition metal compounds. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Ed.; John Wiley & Sons: Chichester, U. K., 1998.
- (18) Kaupp, M. Relativistic effects on NMR chemical shifts. In *Relativistic Electronic Structure Theory*; Schwerdtfeger, P., Ed.; Elsevier: Amsterdam, 2004; Vol. 2.
- (19) Autschbach, J. Calculation of heavy-nucleus chemical shifts: Relativistic all-electron methods. In *Calculation of NMR and EPR Parameters. Theory and Applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004.
- (20) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597–4610.
- (21) van Lenthe, E.; Ehlers, A.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 8943–8953.
- (22) Wolf, A.; Reiher, M.; Hess, B. Transgressing theory boundaries: The generalized Douglas–Kroll transformation. In *Recent Advances in Relativistic Molecular Theory*; Hirao, K., Ishikawa, Y., Eds.; World Scientific: Singapore, 2004; Vol. 5.
- (23) Fukuda, R.; Hada, M.; Nakatsuji, H. *J. Chem. Phys.* **2003**, *118*, 1015–1026.
- (24) Malkin, I.; Malkina, O. L.; Malkin, V. G.; Kaupp, M. *J. Chem. Phys.* **2005**, *123*, 244103–16.
- (25) Visscher, L.; Enevoldsen, T.; Saue, T.; Jensen, H. J. A.; Oddershede, J. *J. Comput. Chem.* **1999**, *20*, 1262–1273.
- (26) Wolff, S. K.; Ziegler, T.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 7689–7698.
- (27) Bouten, R.; Baerends, E. J.; van Lenthe, E.; Visscher, L.; Schreckenbach, G.; Ziegler, T. *J. Phys. Chem. A* **2000**, *104*, 5600–5611.
- (28) Krykunov, M.; Ziegler, T.; van Lenthe, E. *J. Phys. Chem. A* **2009**, *113*, 11495–11500.
- (29) Autschbach, J.; Ziegler, T. *J. Chem. Phys.* **2000**, *113*, 936–947.
- (30) Autschbach, J.; Ziegler, T. *J. Chem. Phys.* **2000**, *113*, 9410–9418.
- (31) Autschbach, J. *J. Chem. Phys.* **2008**, *129*, 094105–9. Erratum *ibid.* **2009**, *130*, 209901.
- (32) Autschbach, J. *ChemPhysChem* **2009**, *10*, 2274–2283.
- (33) Moncho, S.; Autschbach, J. *J. Chem. Theory Comput.* **2010**, *6*, 223–234.
- (34) van Lenthe, E.; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **1997**, *107*, 2488–2498.
- (35) Autschbach, J.; Pritchard, B. *Theor. Chem. Acc.* **2011**, *129*, 453–466.
- (36) Faas, S. *The ZORA approach in ab initio quantum chemistry*, Ph.D. Thesis, Rijksuniversiteit Groningen, The Netherlands, 2000.
- (37) Nichols, P.; Govind, N.; Bylaska, E. J.; de Jong, W. A. *J. Chem. Theory Comput.* **2009**, *5*, 491–499.
- (38) Baerends, E. J.; Ziegler, T.; Autschbach, J.; Bashford, D.; Bérces, A.; Bickelhaupt, F. M.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; van Faassen, M.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; Ghysels, A.; Giammona, A.; van Gisbergen, S. J. A.; Götz, A. W.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Gusarov, S.; Harris, F. E.; van den Hoek, P.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; Kaminski, J. W.; van Kessel, G.; Kootstra, F.; Kovalenko, A.; Krykunov, M. V.; van Lenthe, E.; McCormack, D. A.; Michalak, A.; Mitoraj, M.; Neugebauer, J.; Nicu, V. P.; Noodleman, L.; Osinga, V. P.; Patchkovskii, S.; Philippen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Rodriguez, J. I.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Seldenthuis, J. S.; Seth, M.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visscher, L.; Visser, O.; Wang, F.; Wesolowski, T. A.; van Wezenbeek, E. M.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Yakovlev, A. L. *Amsterdam Density Functional*; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands. URL <http://www.scm.com> (accessed Aug 2011).
- (39) Autschbach, J.; Patchkovskii, S.; Pritchard, B. *J. Chem. Theory Comput.* **2011**, *7*, 2175–2188.
- (40) Dupuis, M. *Comput. Phys. Commun.* **2001**, *134*, 150–166.
- (41) Govind, N.; Valiev, M.; Jensen, L.; Kowalski, K. *J. Phys. Chem. A* **2009**, *113*, 6041.
- (42) Jensen, L.; Govind, N. *J. Phys. Chem. A* **2009**, *113*, 9761.
- (43) Hammond, J.; Govind, N.; Kowalski, K.; Autschbach, J.; Xantheas, S. *J. Chem. Phys.* **2009**, *131*, 214103.
- (44) Andzelm, J.; Rinderspacher, B.; Rawlett, A.; Dougherty, J.; Baer, R.; Govind, N. *J. Chem. Theory Comput.* **2009**, *5*, 2835.
- (45) Kowalski, K.; Krishnamoorthy, S.; Villa, O.; Hammond, J.; Govind, N. *J. Chem. Phys.* **2010**, *132*, 154103.
- (46) Lopata, K.; Govind, N. *J. Chem. Theory Comput.* **2011**, *7*, 1344.
- (47) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (48) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 5151–57.
- (49) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.
- (50) Aquino, F.; Govind, N.; Autschbach, J. *J. Chem. Theory Comput.* **2010**, *6*, 2669–2686.
- (51) Patchkovskii, S.; Ziegler, T. *J. Chem. Phys.* **1999**, *111*, 5730–5740.
- (52) Malkina, O. L.; Vaara, J.; Scimmelpennig, B.; Munzarová, M.; Malkin, V. G.; Kaupp, M. *J. Am. Chem. Soc.* **2000**, *122*, 9206–9218.

- (53) Kaupp, M.; Reviakine, R.; Malkina, O. L.; Arbuznikov, A.; Schimmelpfennig, B.; Malkin, V. G. *J. Comput. Chem.* **2002**, *23*, 794–803.
- (54) Neese, F. *J. Chem. Phys.* **2001**, *115*, 11080–11096.
- (55) Hrobárik, P.; Malkina, O. L.; Malkin, V. G.; Kaupp, M. *Chem. Phys.* **2009**, *356*, 229–235.
- (56) Komorovský, S.; Repiský, M.; Malkina, O. L.; Malkin, V. G.; Malkin, I.; Kaupp, M. *J. Chem. Phys.* **2006**, *124*, 084108–8.
- (57) Repiský, M.; Komorovský, S.; Malkin, E.; Malkina, O. L.; Malkin, V. G. *Chem. Phys. Lett.* **2010**, *488*, 94–97.
- (58) Neese, F. *Coord. Chem. Rev.* **2009**, *253*, 526–563.
- (59) Bühl, M. NMR of transition metal compounds. In *Calculation of NMR and EPR Parameters. Theory and Applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004.
- (60) Straka, M.; Kaupp, M. *Chem. Phys.* **2005**, *311*, 45–56.
- (61) Bühl, M. *Annu. Rep. NMR Spectrosc.* **2008**, *64*, 77–125.
- (62) Schreckenbach, G. *Int. J. Quantum Chem.* **2005**, *101*, 372–380.
- (63) Schreckenbach, G.; Ziegler, T. *J. Phys. Chem. A* **1997**, *101*, 3388–3399.
- (64) Jayatilaka, D. *J. Chem. Phys.* **1998**, *108*, 7587–7594.
- (65) Patchkovskii, S.; Schreckenbach, G. Calculation of EPR g-tensors with density functional theory. In *Calculation of NMR and EPR Parameters. Theory and Applications*; Kaupp, M., Bühl, M., Malkin, V. G., Eds.; Wiley-VCH: Weinheim, Germany, 2004.
- (66) Hrobárik, P.; Repiský, M.; Komorovský, S.; Hrobáriková, V.; Kaupp, M. *Theor. Chem. Acc.* **2011**, *129*, 715–725.
- (67) van Wüllen, C. *J. Chem. Phys.* **1998**, *109*, 392–399.
- (68) Philipsen, P. H. T.; van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *Phys. Rev. B* **1997**, *56*, 13556–13562.
- (69) London, F. *J. Phys. Radium* **1937**, *8*, 397–409.
- (70) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789–807.
- (71) Fukui, H. *Magn. Res. Rev.* **1987**, *11*, 205–274.
- (72) van Lenthe, E. *The ZORA Equation*, Ph.D. Thesis, Vrije Universiteit, Amsterdam, Netherlands, 1996.
- (73) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; van Dam, J. J.; Wang, D.; Apra, E.; Windus, T. L.; Hammond, J.; Autschbach, J.; Aquino, F.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Glaesemann, K.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Vazquez-Mayagoitia, A.; Jensen, L.; Swart, M.; Wu, Q.; Van Voorhis, T.; Auer, A. A.; Noojien, M.; Crosby, L. D.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anshell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsels, H.; Deegan, M.; Dylla, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem*, version 6.0 (2011 developer's version); Pacific Northwest National Laboratory: Richland, WA, 2011.
- (74) Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260–283.
- (75) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Dam, H. J. J. V.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.
- (76) Patchkovskii, S.; Strong, R. T.; Pickard, C. J.; Un, S. *J. Chem. Phys.* **2005**, *122*, 214101–9.
- (77) Vaara, J.; Ruud, K.; Vahtras, O.; Ågren, H.; Jokisaari, J. *J. Chem. Phys.* **1998**, *109*, 1212–1222.
- (78) Vaara, J.; Ruud, K.; Vahtras, O. *J. Chem. Phys.* **1999**, *111*, 2900–2908.
- (79) Roos, B. O.; Lindh, R.; Malmqvist, P.; Veryazov, V.; Widmark, P. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.
- (80) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3295–3305.
- (81) Feller, D. *J. Comput. Chem.* **2000**, *17*, 1571–1586.
- (82) Schuchardt, K.; Didier, B.; Elsethagen, T.; Sun, L.; Gurno-moorthi, V.; Chase, J.; Li, J.; Windus, T. *J. Chem. Inf.* **2007**, *47*, 1045–1052.
- (83) Fukuda, R.; Nakatsuji, H. *J. Chem. Phys.* **2005**, *123*, 044101.
- (84) Hada, M.; Wan, J.; Fukuda, R.; Nakatsuji, H. *J. Comput. Chem.* **2001**, *22*, 1502–1508.
- (85) Ruiz-Morales, Y.; Schreckenbach, G.; Ziegler, T. *J. Phys. Chem. A* **1997**, *101*, 4121–4127.
- (86) Kovacs, A.; Martinsen, K.; Konings, R. *J. Chem. Soc., Dalton Trans.* **1997**, 1037–1042.
- (87) Preut, H.; Wilkes, B.; Naumann, D. *Acta Crystallogr.* **1990**, *C46*, 1113–1115.
- (88) Ziolo, R.; Troup, J. *J. Am. Chem. Soc.* **1983**, *105*, 229–235.
- (89) Ilcyszyn, M.; Lis, T.; Baran, J.; Ratajczak, H. *J. Mol. Struct.* **1992**, *265*, 293–310.
- (90) Gundersen, G.; Hedberg, K. *J. Chem. Phys.* **1978**, *68*, 3548–3552.
- (91) Blom, R.; Haaland, A.; Seip, R. *Acta Chem. Scand.* **1983**, *A37*, 595–599.
- (92) Blake, A.; Pulham, C.; Greene, T.; Downs, A.; Haaland, A.; Verne, H.; Volden, H.; Marsden, C.; Smart, B. *J. Am. Chem. Soc.* **1994**, *116*, 6043–6044.
- (93) Flaud, J.; Arcas, P.; Burger, H.; Polanz, O.; Halonen, L. *J. Mol. Spectrosc.* **1997**, *183*, 310–335.
- (94) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (95) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (96) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (97) Kutzelnigg, W.; Fleischer, U.; Schindler, M. The IGLO-Method: Ab Initio Calculation and Interpretation of NMR Chemical Shifts and Magnetic Susceptibilities. In *NMR Basic Principles and Progress*; Diehl, P., Fluck, E., Gunther, H., Kosfeld, R., Seelig, J., Eds.; Springer-Verlag: Heidelberg, Germany, 1990; Vol. 23.
- (98) Lange, A.; Rohrdanz, M.; Herbert, J. *J. Phys. Chem. B* **2008**, *112*, 6304–6308.
- (99) Belanzoni, P.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **2001**, *114*, 4421–4433.
- (100) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- (101) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (102) Manninen, P.; Lantto, P.; Vaara, J. *J. Chem. Phys.* **2003**, *119*, 2623–2637.
- (103) Manninen, P.; Ruud, K.; Lantto, P.; Vaara, J. *J. Chem. Phys.* **2005**, *122*, 114107. Erratum: *J. Chem. Phys.* **2006**, *124*, 149901.
- (104) Harris, R.; Mann, B. *NMR and the Periodic Table*; Academic Press: London, 1978.
- (105) Jones, C.; Sharma, R.; Naumann, D. *Can. J. Chem.* **1986**, *64*, 987–990.
- (106) Gombler, W. *Naturforsch.* **1981**, *36b*, 535.
- (107) Naumann, D.; Herberg, S. *J. Fluorine Chem.* **1976**, *8*, 1977.
- (108) Drake, J.; Khasrou, L.; Mislankar, A. *Inorg. Chem.* **1994**, *33*, 6154.
- (109) Totsch, W.; Peringer, P.; Sladky, F. *J. Chem. Soc., Chem. Commun.* **1981**, 841.
- (110) Jameson, C. J.; Jameson, A. *Chem. Phys. Lett.* **1987**, *135*, 254–259.
- (111) McFarlane, H.; McFarlane, W. *J. Chem. Soc., Dalton Trans.* **1973**, 2416.
- (112) Gedridge, R. W.; Harris, D. C.; Higa, K. T.; Nissan, R. A. *Organometallics* **1989**, *8*, 2817–2820.
- (113) McFarlane, W.; Wood, R. J. *J. Chem. Soc., Dalton Trans.* **1972**, 1397.
- (114) McFarlane, H.; McFarlane, W. In *Multinuclear NMR* **1987**.
- (115) Jen, C. K.; Foner, S. N.; Cochran, E. L.; Bowers, V. A. *Phys. Rev.* **1958**, *112*, 1169–1182.
- (116) Holmberg, R. W. *J. Chem. Phys.* **1969**, *51*, 3255–3260.
- (117) Van Zee, R. J.; Ferrante, R. F.; Weltner, W., Jr. *J. Chem. Phys.* **1985**, *83*, 6181–6187.
- (118) De Vore, T. C.; Weltner, W. *J. Am. Chem. Soc.* **1977**, *99*, 4700–4703.

- (119) Weltner, W., Jr. *Magnetic Atom and Molecules*; Van Nostrand Reinhold: New York, 1983.
- (120) Knight, L. B., Jr.; Weltner, W., Jr. *J. Chem. Phys.* **1971**, *55*, 2061–2070.
- (121) Knight, L. B., Jr.; Fisher, T. A.; Wise, M. B. *J. Chem. Phys.* **1981**, *74*, 6009–6013.
- (122) Hutchison, C. A., Jr.; Weinstock, B. *J. Chem. Phys.* **1960**, *32*, 56–61.
- (123) Dupuis, M.; Rys, J.; King, H. *J. Chem. Phys.* **1976**, *65*, 111–116.
- (124) Rys, J.; Dupuis, M.; King, H. *J. Comput. Chem.* **1983**, *4*, 154–157.

Theoretical Study of the Luminescent States and Electronic Spectra of UO_2Cl_2 in an Argon Matrix

Jing Su,[†] Yi-Lei Wang,[†] Fan Wei,[†] W.H.E. Schwarz,^{†,‡} and Jun Li^{*,†}

[†]Department of Chemistry and Laboratory of Organic Optoelectronics and Molecular Engineering of the Ministry of Education, Tsinghua University, Beijing 100084, China

ABSTRACT: The electronic absorption and emission spectra of free UO_2Cl_2 and its Ar-coordinated complexes below $27\,000\text{ cm}^{-1}$ are investigated at the levels of ab initio complete active space second-order perturbation theory (CASPT2) and coupled-cluster singles and doubles and perturbative triples [CCSD(T)] using valence 3ζ -polarized basis sets. The influence of the argon matrix in the 12K experiment on the electronic spectra is explored by investigating the excited states of argon complexes $\text{Ar}_n\text{UO}_2\text{Cl}_2$. The calculated two most stable complexes with $n = 2, 3$ can explain the observed two matrix sites corresponding to the experimental two-component luminescence decay. In these uranyl complexes, Ar-coordination is found to have little influence on the $^3\Phi$ ($\Omega = 2_g$) character of the luminescent state and on the electronic spectral shape. The calculations yield a coherent assignment of the experimental excitation spectra that improves on previous assignments. The simulated luminescence spectral curves based on the calculated spectral parameters of UO_2Cl_2 from both CASPT2 and CCSD(T) agree well with experiment.

INTRODUCTION

The study of optical properties of uranyl (UO_2^{2+}) compounds has a long history.^{1–3} Various systems exhibit related absorption and emission spectra, in particular in the low-temperature region. The fluorescence spectra are characterized by several vibrational progressions based on a common electronic origin. The intensity distribution of the fluorescence spectra changes with the ligands coordinated to the uranyl moiety. That is, the spectral shapes inform about the electronic and geometric structures of the respective uranyl complexes.

Time-resolved laser-induced fluorescence (TRLIF) has become a common tool to study the speciation of actinides (IUPAC: actinoids) in natural and artificial environments.⁴ This spectroscopic technique is important for an atomistic understanding of the interactions of actinides with various inorganic and organic/biochemical ligands and provides the basic information for handling actinide contaminations in the environment and in biological systems, including the human body.^{5,6} Theory-assisted reconstruction of knowledge from experimental data is nowadays possible for both weakly perturbed molecules or building groups as well as for complex condensed phases in thermodynamic equilibrium.⁷

We have recently investigated structures, stabilities, and the vibration-resolved luminescence of uranyl–glycine–water complexes in solution, applying comparatively simple theoretical strategies.⁷ To improve on the level of spectral simulations, advanced correlated ab initio methods including spin–orbit (SO) coupling are required for open-shell excited states, refining simple density functional theory (DFT) which is often applicable for actinide closed-shell ground-state species. However, SO-complete active space second-order perturbation theory (CASPT2) or SO-multireference configuration interaction with single and double substitutions (MRCISD) are extremely demanding for such systems, due to the complexity of the actinide valence shell and the diversity and lack of symmetry of the

coordination structures embedded in first and second solvent shells. On the other hand, many details of the experimental solution spectra are washed out, in particular concerning the rotational, vibrational, and electronic near-degenerate fine structure.

Recently, Heaven's group has reported electronic spectra of single UO_2Cl_2 molecules in a solid Ar matrix,⁸ where UO_2Cl_2 has a significantly simpler coordination structure than the multitude of uranyl complexes in aqueous solution. The peaks of the fluorescence spectra remain sharp and well-separated; the assignment of different vibrational modes and low-electronic states comes into reach. Accordingly, the $\text{Ar}_n\text{—UO}_2\text{Cl}_2$ system is a promising candidate for the high-level path of theoretical calculations and spectra simulations.

The purpose of the present paper is the theoretical analysis of geometric and electronic structure and the computational simulation of luminescence of UO_2Cl_2 by using state-of-the-art quantum chemical methods. Before describing the theoretical methodology (Computational Details) and analyzing the computational results (Computational Results and Discussion), we briefly summarize the current understanding of uranyl and the state of experimental and theoretical knowledge of the uranyl chloride complexes.

Model Concepts for Uranyl Compounds. In order to understand chemical bonding, excited states, and absorption and emission spectra of uranyl and its complexes (Figure 1), we refer to the axial-symmetric molecular orbitals (MOs) of bare uranyl in Figure 2.⁹ At the nonrelativistic and scalar relativistic levels of approximation, the U-5f type MOs under $D_{\infty h}$ symmetry are split into σ_u and pairs of π_w , δ_u and ϕ_u and the U-6d type MOs into σ_g and pairs of π_g and δ_g . In an ionic picture, all valence electrons are assigned to O^{2-} , and uranium obtains the oxidation state U^{VI} . The U-7s shell is energetically pushed up upon ligand coordination on

Received: June 19, 2011

Published: August 25, 2011

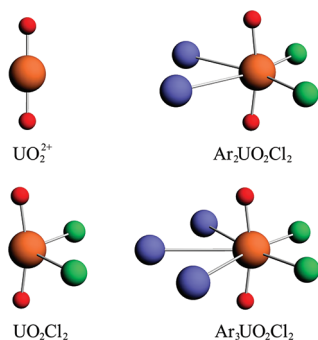


Figure 1. Geometric Structures of UO_2^{2+} , UO_2Cl_2 , $\text{Ar}_2\text{UO}_2\text{Cl}_2$, and $\text{Ar}_3\text{UO}_2\text{Cl}_2$.

U(VI) and can be neglected in qualitative discussions of uranyl complexes.

The occupied $2p\sigma, \pi$ valence shells of the two terminal O^{2-} anions are stabilized by the U- $5f, 6d$ manifold of valence orbitals with matching symmetry. The 12 valence electrons of UO_2^{2+} participate in dative bonding through σ_g and σ_u and pairs of π_u and π_g canonical MOs, corresponding to $\text{U}\equiv\text{O}$ triple bonding. Herein σ_u is the highest occupied molecular orbital (HOMO) due to a push-from-below via U- $6p$ mixing, while the order of the slightly lower, near-degenerate σ_g, π_w and π_g levels varies somewhat with equatorial coordination. The antibonding counterparts remain empty, see Figure 2.

In the bare uranyl dication, the U- $5f\delta_w, 5f\phi_u$ and $6d\delta_g$ type orbitals have no symmetry-matching counterparts from the O atoms and do not participate in O–U–O bonding. These empty orbitals remain nonbonding and localized on U but are available for weaker σ and π donor interactions of Lewis bases in the equatorial plane. Thus, the ground states of the uranyl–dichloride and dichloride–argon complexes are of closed-shell type and derive from the $^1\Sigma_g^+$ state in $D_{\infty h}$ symmetry. The lowest excited states correspond to electronic transitions from U–O bonding σ_u to nonbonding U- $5f\phi_u$ and $5f\delta_u$ type MOs (see the arrow in Figure 2), yielding $^1,3\Delta_g$ and $^1,3\Phi_g$ states. With the inclusion of SO-coupling, $^3\Delta_g$ splits into Π_g, Δ_g and Φ_g ($\Omega = 1_g, 2_g, 3_g$) and $^3\Phi_g$ splits into Δ_g, Φ_g and Γ_g ($\Omega = 2_g, 3_g, 4_g$).

Known Uranyl Chloride Complexes. Experimentally, the absorption and fluorescence spectra of $[\text{UO}_2\text{Cl}_4]^{2-}$ in crystals have been extensively investigated by Denning and others.^{10,11} Due to environmental concerns of uranyl ions released from nuclear waste, much work has been invested on the solution chemistry of uranyl complexes. Görller-Walrand et al.¹² reported that uranyl chloride complexes in various organic solvents have spectra consistent with those from crystalline materials.

Theoretical investigations of excited states of actinide compounds have been challenging. Early efforts include calculations with various DFT approximations.^{2,13,14} Zhang, Matsika, and Pitzer^{15,16} started the ab initio theoretical investigation of electronic excited states of UO_2^{2+} and $\text{Cs}_2\text{UO}_2\text{Cl}_4$ with the SO–Cl approach. Afterward, Pierloot et al.¹⁷ investigated $[\text{UO}_2\text{Cl}_4]^{2-}$ applying SO-CASPT2 methods, obtaining better agreement for the experimental electronic excitations and the O–U–O symmetric stretching vibration. They also performed time-dependent (TD)DFT calculations with state-averaged optimized potentials (SAOP) and SO-coupling and obtained smaller U–O bond length expansions for the excited states and a different luminescent state than with CASPT2.¹⁸ The

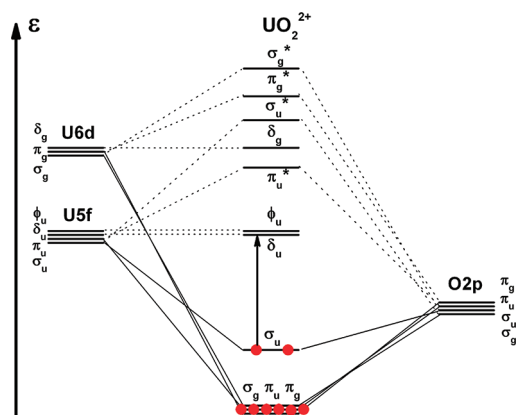


Figure 2. Qualitative scalar-relativistic valence orbital energy level schemes of U and O atoms and uranyl ($D_{\infty h}$) in the middle. The vertical arrow indicates the lowest electronic excitations.

influence of solvent coordination in acetone on the absorption spectra of UO_2Cl_2 and UO_2Cl_3^- has also been studied with the CASPT2 method.¹⁹

COMPUTATIONAL DETAILS

Structures and spectra of free UO_2Cl_2 and its Ar-coordinated complexes were investigated by using CASPT2 and CCSD(T) methods, with and without SO-coupling, as implemented in the MOLPRO 2008.1 program.²⁰ Various $\text{Ar}_n\text{UO}_2\text{Cl}_2$ complexes were treated in a first step at the DFT level, using the ADF 2009.1 software.²¹

It has become well-known that atomic effective core potentials (ECPs) can reduce the computational expenses of self-consistent field (SCF), multiconfiguration self-consistent field (MCSCF) and CC approaches drastically. When using ECP-adapted basis sets together with ECPs, there is no loss of reliability as long as the core–valence sets are appropriately selected.²² In addition, relativistic ECPs (RECPs) inherently account for SO-coupling which also holds for the relativistic zeroth-order regular approximation (ZORA). Therefore, we used the Stuttgart energy-consistent RECPs²³ for Cl, Ar, and U, where the $1s^2 2s^2 2p^6$ cores of Cl and Ar were treated by the scalar ECP10MWB ones, while optimizing the $3s3p$ valence shells, and the $1s^2 - 4f^{14}$ core for U was treated by the scalar and SO-coupled ECP60MWB one, with the $5spdf, 6spd,$ and $7sp$ semicore and valence shells optimized. We applied the 6-311+G* basis set for O,²⁴ the ECP10MWB for Cl and Ar with an additional d-polarization function ($\zeta=0.75$) for Cl, and ECP60MWB-SEG basis set for U, respectively. The atomic core shells and the U- $5spdf$ were not correlated.

In the DFT calculations, we used the PBE functional and the scalar relativistic ZORA approach. The frozen core approximation and TZ2P bases were applied.²⁵

Geometries and Vibrational Frequencies. Geometric optimizations of the electronic ground states of UO_2^{2+} in $D_{\infty h}$ and of UO_2Cl_2 in C_{2v} symmetry with CASPT2 and CCSD(T) were converged to gradients $<1.0 \times 10^{-4}$. The most stable $\text{Ar}_n\text{UO}_2\text{Cl}_2$ complexes were at first screened out for $n = 1-4$, using the ADF binding energy analysis and were then further optimized with CCSD(T) calculations. The two most stable structures were also optimized with the CASPT2 approach. Vibrational analyses were performed for the UO_2Cl_2 ground state at the CCSD(T) level.

Table 1. Correlation of Symmetry Species of Point Groups $D_{\infty h}$ and C_{2v}

$D_{\infty h}$	C_{2v}
Σ_g^+	A_1
Σ_u^+	B_1
Π_g, Δ_w, Φ_g	$A_2 + B_1$
Π_w, Δ_g, Φ_u	$A_1 + B_2$

Born–Oppenheimer (BO) potential energy curves of the SO-averaged and SO-coupled excited electronic states versus the U–O distances were at first scanned in steps of 1 pm, with the other geometric parameters fixed at their ground-state values. The expansions of the U–O distances in the excited states were obtained from polynomial interpolation. For the lowest excited luminescent state, the equilibrium values of the other geometric parameters were then similarly approximated, keeping the U–O distance of the state fixed. Thereby, the approximate U–O equilibrium distances, vertical and adiabatic excitation energies, and O–U–O symmetric stretching frequencies were determined. The error of this approximation applied to the ground state was less than 2 cm^{-1} .

Electronic States. For simplicity, we use approximate $D_{\infty h}$ symmetry notations for orbitals and states of all species, except where explicitly noted otherwise. The relations between $D_{\infty h}$ and C_{2v} (UO_2Cl_2 , $\text{Ar}_{2,3}\text{UO}_2\text{Cl}_2$) symmetry species within the coordinate system used in our calculations are given in Table 1.

RASSCF/CASPT2/SO Calculations. The active spaces for ground-state CASSCF calculations of all molecular species were confined to the UO_2^{2+} moiety: The six bonding and six antibonding (*) MOs of $\sigma_g, \sigma_w, \pi_g,$ and π_u type from the U-5f,6d and two O-2p shells with 12 valence electrons were correlated by CAS-(12,12). The active spaces for the excited states contained in addition nonbonding U-5f type orbitals of δ_u or ϕ_u symmetry (Figure 2), giving 12 electrons in 14 orbitals for $D_{\infty h}\text{-UO}_2^{2+}$, denoted as CAS(12,14) hereafter, or 12 electrons in 13 orbitals for $C_{2v}\text{-UO}_2\text{Cl}_2$ and $C_{2v}\text{-Ar}_{2,3}\text{UO}_2\text{Cl}_2$, i.e., CAS(12,13). In the equatorially ligated uranyl species (Figure 1), the degeneracy of δ_u and ϕ_u is lifted, with little orbital and configuration mixing, as known from the literature.^{17,26} CAS(12,16) calculations with both δ_u or ϕ_u pairs simultaneously in the active space were deemed unnecessary.

SO-averaged CASPT2 calculations were performed on the ground states and on all excited states arising from single excitations out of the σ_u HOMO into the nonbonding orbitals of U-5f δ_u, ϕ_u type, which gives four singlets and four triplets. Individually optimized CASSCF orbitals were used for each state, except for the singlet excited states of same symmetry as the ground state. Here, the ground-state orbitals helped converging to correct occupation schemes. A level shift of 0.3 au was applied to improve the CASPT2 convergence.

For UO_2^{2+} , the g_1 -corrected CAS Fock-operators were also employed in order to obtain a balanced treatment of closed-shell ground and open-shell excited states. Because of the near degeneracy of some excited states of UO_2Cl_2 , this was not always feasible. Therefore, the g_1 -corrections of respective states of UO_2^{2+} were added to the uncorrected values of the states of UO_2Cl_2 , as suggested by Pierloot (designated as g_1').¹⁷

SO coupling was treated by a restricted RAS-SI/SO approach^{27,28} in an active space of the mentioned 16 orbitals, labeled RAS(12,16), where up to 4 electrons were excited. The SO-averaged restricted active space self-consistent field (RASSCF)

Table 2. Geometric Parameters of UO_2^{2+} and UO_2Cl_2 from CASPT2 and CCSD(T) Calculations for the Ground and Four Low Excited Triplet States at the SO-Averaged Level

molecule	state in				
	$C_{2v}(D_{\infty h})$	$R_{\text{U-O}}$, pm	$R_{\text{U-Cl}}$, pm	$\angle \text{OUO}$, °	$\angle \text{ClUCl}$, °
CASPT2					
UO_2^{2+}	$(^1\Sigma_g^+)$	169.37		180.0	
UO_2Cl_2	$^1A_1(^1\Sigma_g^+)$	175.29	249.12	165.8	106.8
CCSD(T)					
UO_2^{2+}	$(^1\Sigma_g^+)$	169.03		180.0	
UO_2Cl_2	$^1A_1(^1\Sigma_g^+)$	174.77	250.92	165.8	107.7
	$^3A_1(^3\Delta_g)$	179.73	250.19	161.8	102.8
	$^3B_2(^3\Delta_g)$	179.45	251.65	164.5	104.2
	$^3B_1(^3\Phi_g)$	181.18	251.10	158.2	101.5
	$^3A_2(^3\Phi_g)$	181.08	252.09	158.6	102.3

singlet and triplet wave functions were determined in the basis of state averaged (SA) RASSCF orbitals of the ground and all excited singlet states. The resulting RASSCF wave functions were then used to construct a 17×17 SO coupling matrix, where the diagonal elements were correlation corrected by using the CASPT2 energies. Single-point SA-CAS(12,16)-SCF test calculations confirmed that the errors remain $<10 \text{ cm}^{-1}$ for UO_2^{2+} and $<20 \text{ cm}^{-1}$ for UO_2Cl_2 . This combination of RASSCF/SI-SO with CASPT2 energy is labeled as RASSCF/CASPT2/SO. Calculations on UO_2^{2+} , NUO^+ , UN_2 , and UF_6 molecules using this approach have well reproduced the experimental data.^{26,29}

RASSCF/CCSD(T)/SO Calculations. The CASSCF/CCSD-(T)/SO method had been shown to be quite accurate for various properties of transition-metal compounds.^{30–32} We here applied such an approach to the excited states of UO_2Cl_2 and $\text{Ar}_{2,3}\text{UO}_2\text{Cl}_2$, but constructed the SO coupling matrix from SA-RASSCF wave functions. For the ground and excited triplet state energy curves, the SO-averaged scheme was applied at first. Then the CASPT2 energy difference between the excited singlet and triplet pair was obtained pointwise to estimate the excited singlet energy curves. The SO coupling effect was included in the same way as above with the diagonal elements corrected by the CCSD(T) state energies.

Simulation of Luminescent Spectra and Normal Coordinates Analysis. Following our previous work,⁷ the profiles of the luminescence spectra were modeled using the Franck–Condon formulas of Fonger and Struck.³³ In this approach vibrational frequency changes upon electronic transition are considered, while neglecting anharmonicities and Duschinsky rotations, which are expected to be small for the cases at hand. Ground-state geometry optimizations and frequency calculations of UO_2Cl_2 were also performed with DFT/PBE using Gaussian 03 with the same basis sets as in the CASPT2 and CCSD(T) calculations above.³⁴

The obtained Cartesian force constant matrix was used to construct the F matrix of the Wilson–Decius FG method³⁵ for the chosen set of internal coordinates and then combined with the G matrix to achieve a normal coordinate analysis. The F and G matrices were obtained with the help of the program of McIntosh and Peterson.^{36,37} The dimensionless geometric displacement parameter upon electronic transition for O–U–O, Δ and the Huang–Rhys factor, S ,³⁸ which characterize the overall

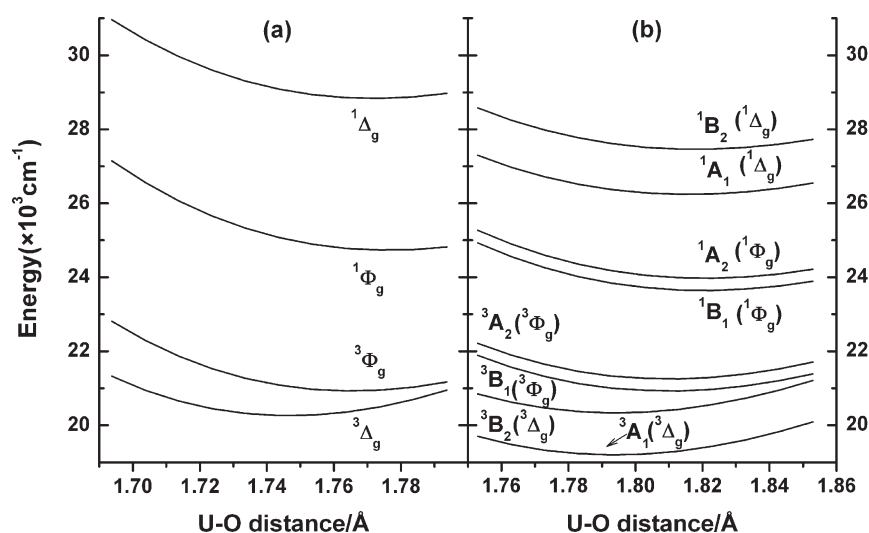


Figure 3. Energy curves of the low-lying excited states of (a) UO_2^{2+} and (b) UO_2Cl_2 for the O–U–O symmetric stretch at the scalar relativistic level from CASPT2 calculations without g_1 correction.

Table 3. Spectroscopic Data from SO-Averaged CASPT2-[g_1] for UO_2^{2+} (g_1 correction in parentheses) and UO_2Cl_2 (g_1' corrected)

molecule	$C_{2v}(D_{\infty h})$	E^a , cm^{-1}	R_e^b , pm	T_e^c , cm^{-1}	ν_s^d , cm^{-1}
UO_2^{2+}	$X^1\Sigma_g^+$		169.35 (0.02)	0 (0)	993 (0)
	$^3\Delta_g$	22 445 (1118)	175.16 (0.32)	21 229 (976)	865 (5)
	$^3\Phi_g$	23 954 (1143)	176.92 (0.36)	21 843 (934)	879 (5)
	$^1\Phi_g$	28 295 (1148)	177.96 (0.46)	25 566 (851)	879 (3)
	$^1\Delta_g$	31 826 (862)	177.59 (0.47)	29 387 (572)	868 (5)
UO_2Cl_2	$X^1A_1(X^1\Sigma_g^+)$		175.27		865
	$a^3A_1(^3\Delta_g)$	20 816	179.82	20 182	757
	$a^3B_2(^3\Delta_g)$	21 962	179.87	21 317	758
	$a^3B_1(^3\Phi_g)$	23 029	181.49	21 855	771
	$a^3A_2(^3\Phi_g)$	23 359	181.51	22 178	771
	$a^1B_1(^1\Phi_g)$	26 075	182.58	24 482	764
	$a^1A_2(^1\Phi_g)$	26 418	182.61	24 809	765
	$a^1A_1(^1\Delta_g)$	28 157	182.16	26 803	739
	$a^1B_2(^1\Delta_g)$	29 441	182.37	28 019	739

^aVertical excitation energy E at ground-state geometry. ^bU–O equilibrium distance R_e . ^cCalculated adiabatic excitation energy T_e . ^dO–U–O symmetric stretching frequency ν_s .

shape of the vibrational intensity distribution, were finally obtained from eqs 1–3:

$$FGL = L\Lambda, \quad L^T F^{-1} L = \Lambda^{-1} \quad (1)$$

$$Q = L^T R, \quad \Delta Q = L^T \Delta R \quad (2)$$

$$\Delta = \sqrt{(\omega/\hbar)} \cdot \Delta Q, \quad S = \Delta^2/2 \quad (3)$$

COMPUTATIONAL RESULTS AND DISCUSSION

Optimized Structures of UO_2^{2+} and UO_2Cl_2 . As shown in Table 2, the optimized ground-state structures of UO_2^{2+} and UO_2Cl_2 from the CASPT2 and CCSD(T) methods are consistent, both indicating multiple U–O and dative U–Cl bonding.

CCSD(T) gives ~ 0.4 pm shorter U–O distances but ~ 0.2 pm longer U–Cl distances than CASPT2. The expansion of the U–O bonds of uranyl upon equatorial coordination (~ 6 pm for two Cl^- ligands) is well understood as equatorial orbital interaction slightly weakening the U–O triple bond in UO_2Cl_2 .³⁹ The bending of the linear OUO unit by $\sim 14^\circ$ can be rationalized in terms of 5f-6d-7s hybridization, electrostatic Cl^- - O^{2-} repulsion or valence-shell electron-pair repulsion.

The lowest excited states of UO_2Cl_2 are dominated by $\sigma_u \rightarrow \delta_u$ or ϕ_u transitions (Figure 2), similar to those in many other uranyl complexes.⁷ The U–O distances of the triplet SO-averaged states are expanded by 4.8 or 6.4 pm, respectively, while the U–Cl distances vary by no more than 1.2 pm. The OUO and ClUCl angles are reduced by 1.3 up to 7.5°.

SO-Averaged Potential Energy Curves of Excited States. CASPT2 Results. CASPT2 energy curves of the lowest states of

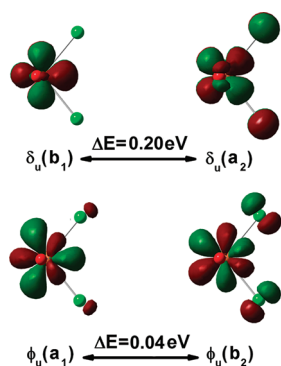


Figure 4. MO isosurfaces (isovalue = 0.03 au) and energies of U-Sf δ_u and ϕ_u orbitals of UO_2Cl_2 from DFT/PBE calculations.

Table 4. Spectroscopic Data of UO_2Cl_2 from SO-Averaged CCSD(T) Scan Calculations^a

state	E , cm^{-1}	R_{eq} , pm	T_{e} , cm^{-1}	ν_{s} , cm^{-1}
X^1A_1		174.77		901
a^3A_1	21 389	179.60 (179.73)	20 710 (20 573)	772
a^3B_2	22 476	179.58 (179.45)	21 803 (21 764)	773
a^3B_1	23 161	180.98 (181.18)	22 003 (21 702)	789
a^3A_2	23 514	181.01 (181.08)	22 343 (22 056)	790
a^1B_1	26 244	181.92	24 731	785
a^1A_2	26 613	181.95	25 075	788
a^1A_1	29 062	181.71	27 744	755
a^1B_2	30 287	181.81	28 921	756

^a See footnotes of Table 3. CCSD(T) optimized results in parentheses.

UO_2^{2+} and UO_2Cl_2 arising from the $\sigma_u \rightarrow \delta_u$, ϕ_u excitations at the SO-averaged level are displayed in Figure 3. Respective g_1 and g_1' -corrected numerical spectroscopic data are collected in Table 3. The well-known singlet–triplet splittings of UO_2^{2+} , 9.4 and $4.3 \times 10^3 \text{ cm}^{-1}$ for the Δ_g and Φ_g , respectively,^{17,40} are nearly carried over to UO_2Cl_2 (7.4 and $3.1 \times 10^3 \text{ cm}^{-1}$). The chloride ligand-field splittings are 1.2 and $0.3 \times 10^3 \text{ cm}^{-1}$. The large $\Delta-\Phi$ difference corresponds to the angular momenta, the shapes, and the orbital energy differences of the U-Sf δ_u and U-Sf ϕ_u type MOs, displayed in Figure 4. Remarkably, there is no combination of Cl-2p π with U-Sf δ_u of B_1 symmetry, also found by Pierloot¹⁷ for the similar case of $\text{UO}_2\text{Cl}_4^{2-}$, while the Cl-2p σ mixes more strongly with U-Sf $\delta_u(A_2)$ than with U-Sf $\phi_u(A_1)$.

The O–U–O distances in UO_2Cl_2 are several pm longer than in UO_2^{2+} ; correspondingly, the symmetric stretching frequencies $\nu_s(\text{OUO})$ are smaller, by 128 cm^{-1} for the ground states and by around 115 cm^{-1} for the excited states. The ν_s values of the excited states are smaller than those of the ground states throughout by more than 100 cm^{-1} , corresponding to the U–O bond length expansion upon electronic excitation. The adiabatic excitation energies of UO_2^{2+} and UO_2Cl_2 differ in the range of -2584 to $+335 \text{ cm}^{-1}$.

CCSD(T) Results. Spectral parameters of UO_2Cl_2 from SO-averaged CCSD(T) scans and then fully optimized CCSD(T)

are displayed in Table 4. The former reproduces the latter for the excited triplet states quite well, within 2 pm and 300 cm^{-1} . The CCSD(T) scans of the BO energy curves yield reasonable approximations to the CCSD(T) optimized results. We expect similarly reliable results from the CASPT2 scans.

The U–O bond lengths of the ground and excited states from CCSD(T) in Table 4 are around 0.5 pm shorter than the CASPT2- $[g_1']$ results in Table 3. The excitation energies in Table 4 are larger than the CASPT2 values by 150 to 950 cm^{-1} , and the ν_s values are larger: 38 cm^{-1} for the ground state, 20 cm^{-1} for the excited Φ_g type states, and 16 cm^{-1} for the excited Δ_g type states. These results illustrate the accuracy of the theoretical results.

Spin–Orbit Coupled States. *RASSCF/CASPT2/SO Results.* SO-coupled BO-energy curves of the low-lying excited states of UO_2^{2+} and UO_2Cl_2 along the O–U–O symmetric stretch are presented in Figure 5. These curves are derived from the spin triplets and are calculated with CASPT2. The g_1 and g_1' -corrected numerical data with Ω -values are collected in Table 5. The SO-splittings of the states of $^3\Delta_g$ type are rather regular and of the order of $2 \times 10^3 \text{ cm}^{-1}$. The SO splittings of the $^3\Phi_g$ type states are larger due to the larger orbital angular momentum component (nearly $6 \times 10^3 \text{ cm}^{-1}$) and less regular due to more pronounced interaction of e–e and SO configuration mixing. As a result, the lowest excited state is dominated by $^3\Phi_g$ character.^{7,17,18,40a}

Literature results on the lowest excited, luminescent state of UO_2^{2+} are also displayed in Table 5. Our results agree better with those of Pierloot and van Besien¹⁷ than with those of Zhang and Pitzer¹⁵ or Réal et al.,⁴⁰ concerning the $^3\Phi_g$ character, the U–O bond length expansion and the excitation energy. From the experimental adiabatic excitation energy of $\sim 20\,323 \text{ cm}^{-1}$ and the deduced U–O bond length expansion of the luminescent state of ~ 5.8 pm and the respective errors of the CASPT2 calculation of UO_2Cl_2 , the most probable values for UO_2^{2+} are deduced to be $20\,434 \text{ cm}^{-1}$ and 7.5 pm, consistent with the best excitation energy given by Vallet^{40a} and the best U–O bond length expansion by Pierloot.¹⁷ As usual,^{40b} the calculated luminescence parameters are influenced by the different basis sets and the treatment of electron correlation and relativistic effects.

While there is no excited-state calculations on UO_2Cl_2 , theoretical results on related $\text{UO}_2\text{Cl}_2(\text{ac})_3$ (ac = acetone) are known.¹⁹ Inasmuch as stronger ligand interactions in the equatorial plane destabilize $^3\Phi_g$ type states in comparison to $^3\Delta_g$ ones,^{7,17,41} the lowest excited state of $\text{UO}_2\text{Cl}_2(\text{ac})_3$ was found to be ($^3\Delta_g$) aA_2 with a vertical energy around $20.3 \times 10^3 \text{ cm}^{-1}$, i.e., higher than for UO_2Cl_2 .

RASSCF/CCSD(T)/SO Results. RASSCF/CCSD(T)/SO results for UO_2Cl_2 are displayed in Table 6 and are similar to the RASSCF/CASPT2- $[g_1']$ /SO ones. Comparing with the CASPT2 results, the CCSD(T) approach yields up to 0.6 pm smaller U–O distances, 160 to 525 cm^{-1} higher excitation energies, and 5 to 30 cm^{-1} larger vibration frequencies. These differences are also found at the SO-averaged level (Tables 3 and 4).

Influence of the Ar Matrix. *Structures, Binding Energies, and Excited States of $\text{Ar}_n\text{UO}_2\text{Cl}_2$.* As weak Lewis bases, argon atoms can bind to actinide metal atoms at lower temperatures. For instance, the coordination of Ar atoms to CUO and UO_2 has been shown to cause so-called ground-state reversal in comparison to gas-phase or neon matrix results.^{42–46}

Since Ar is a much weaker Lewis base than carbonyl or carboxyl ligands, oxophilic uranyl is expected to be only weakly

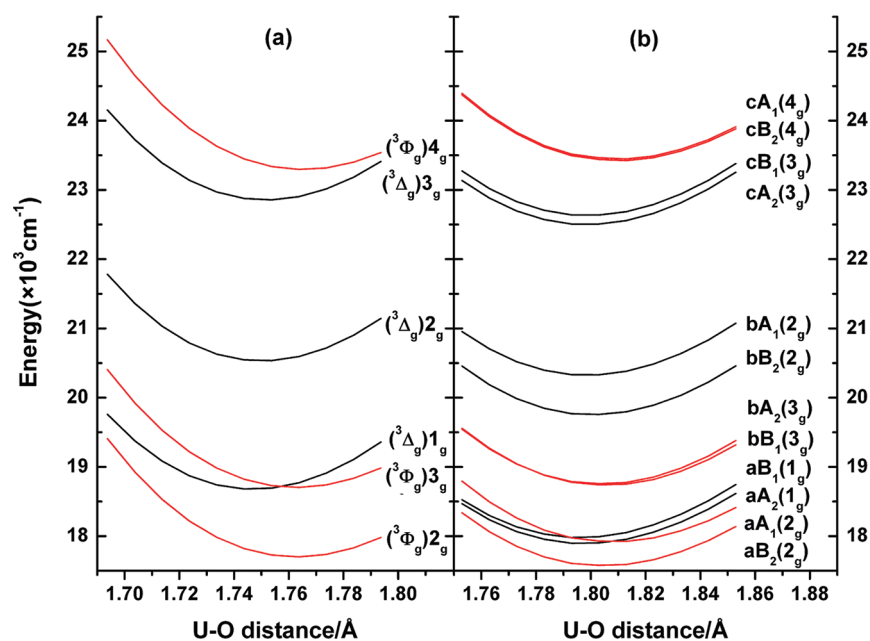


Figure 5. Energy curves of excited states of (a) UO_2^{2+} and (b) UO_2Cl_2 along the O–U–O symmetric stretch coordinate from CASPT2 with SO coupling, without g_1 -correction (red curves: ${}^3\Phi$ type; black curves: ${}^3\Delta$ type).

Table 5. Spectroscopic Data from SO-Coupled CASPT2- $[g_1]$ for UO_2^{2+} (g_1 correction in parentheses) and UO_2Cl_2 (g_1' corrected)^a

molecule	state (Ω)	main ^b	E , cm^{-1}	$R_{e'}$, pm	$\Delta R_{e'}$, pm ^c	T_e , cm^{-1}	ν_s , cm^{-1}
UO_2^{2+}	$X\Sigma_g^+(0_g^+)$			169.35 (0.02)		0 (0)	993 (0)
	$a\Delta_g(2_g)$	${}^3\Phi_g$	20 543 (1134)	176.71 (0.37)	+7.36	18 620 (937)	862 (3)
	$\Pi_g(1_g)$	${}^3\Delta_g$	20 877 (1118)	175.19 (0.32)	+5.84	19 648 (975)	866 (6)
	$a\Phi_g(3_g)$	${}^3\Phi_g$	21 541 (1135)	176.71 (0.37)	+7.36	19 618 (932)	862 (4)
	$b\Delta_g(2_g)$	${}^3\Delta_g$	22 892 (1110)	175.49 (0.32)	+6.14	21 471 (950)	886 (6)
	$b\Phi_g(3_g)$	${}^3\Delta_g$	25 282 (1127)	175.65 (0.33)	+6.30	23 809 (966)	880 (6)
	$\Gamma_g(4_g)$	${}^3\Phi_g$	26 312 (1144)	176.90 (0.35)	+7.55	24 211 (935)	878 (4)
Pierloot ¹⁷	$a\Delta_g(2_g)$	${}^3\Phi_g$	19 195		7.4	17 227	815
Vallet ^{40a}	$a\Delta_g(2_g)$	${}^3\Phi_g$	22 789		6.4	21 338	963
Pitzer ¹⁵	$\Pi_g(1_g)$	${}^3\Delta_g$			6.5	20 719	867
Vallet ^{40b}	$\Pi_g(1_g)$	${}^3\Delta_g$	18 610		4.1	17 557	
UO_2Cl_2	$XA_1(0_g)$			175.27			865
	$aB_2(2_g)$	${}^3\Phi_g$	19 471	180.97	+5.70	18 509	752
	$aA_1(2_g)$	${}^3\Phi_g$	19 928	181.30	+6.03	18 850	760
	$aA_2(1_g)$	${}^3\Delta_g$	19 587	180.25	+4.98	18 874	741
	$aB_1(1_g)$	${}^3\Delta_g$	19 644	180.11	+4.84	18 961	743
	$bB_1(3_g)$	${}^3\Phi_g$	20 693	181.04	+5.77	19 666	772
	$bA_2(3_g)$	${}^3\Phi_g$	20 681	180.90	+5.63	19 686	776
	$bB_2(2_g)$	${}^3\Delta_g$	21 567	180.55	+5.28	20 706	775
	$bA_1(2_g)$	${}^3\Delta_g$	22 065	180.33	+5.06	21 278	766
	$cA_2(3_g)$	${}^3\Delta_g$	24 262	180.34	+5.07	23 470	768
	$cB_1(3_g)$	${}^3\Delta_g$	24 401	180.37	+5.10	23 602	769
	$cB_2(4_g)$	${}^3\Phi_g$	25 519	181.46	+6.19	24 351	769
	$cA_1(4_g)$	${}^3\Phi_g$	25 538	181.45	+6.18	24 374	770
	$\text{UO}_2\text{Cl}_2\text{ac}_3$ ^d	$aA_2(1_g)$	${}^3\Delta_g$	20 338			

^a See footnotes of Table 3. ^b Main component of wave function. ^c U–O bond length expansion upon electronic excitation. ^d See ref 19, ac = acetone.

perturbed by Ar. We simulate the argon matrix effect by $\text{Ar}_n\text{UO}_2\text{Cl}_2$ complexes, neglecting additional soft lattice effects.

Structures of $\text{Ar}_n\text{UO}_2\text{Cl}_2$ molecules ($n = 0-4$) exhibiting C_{2v} symmetry were optimized by using DFT/PBE and CCSD(T)

with basis set superposition error (BSSE) correction. From the binding energies shown in Figure 6, the most stable species are $\text{Ar}_2\text{UO}_2\text{Cl}_2$ and $\text{Ar}_3\text{UO}_2\text{Cl}_2$. These species were chosen to simulate the Ar matrix effects on the UO_2Cl_2 spectra and their structural parameters from CCSD(T) and CASPT2 calculations are displayed in Table 7.

The binding energy per Ar atom is 10 and 5.5 kJ/mol in $\text{Ar}_2\text{UO}_2\text{Cl}_2$ and $\text{Ar}_3\text{UO}_2\text{Cl}_2$, respectively. The U–Ar distances of 3.1–3.4 Å are consistent with van der Waals binding. Equatorial coordination of Ar atoms expands the U–O distances, albeit only by 0.1 pm per Ar. The O–U–O bending due to the Cl ligands is

Table 6. Spectroscopic data of UO_2Cl_2 from RASSCF/CCSD(T)/SO^a

state (Ω)	main	E , cm^{-1}	R_e , pm	ΔR_e , pm	T_e , cm^{-1}	ν_s , cm^{-1}
$\text{XA}_1(0_g)$			174.77			901
$\text{aB}_2(2_g)$	$^3\Phi_g$	19 804	180.66	+5.89	18 797	775
$\text{aA}_1(2_g)$	$^3\Phi_g$	20 157	180.87	+6.10	19 060	782
$\text{aA}_2(1_g)$	$^3\Delta_g$	20 106	180.20	+5.43	19 297	752
$\text{aB}_1(1_g)$	$^3\Delta_g$	20 183	180.05	+5.28	19 422	748
$\text{bB}_1(3_g)$	$^3\Phi_g$	20 991	180.47	+5.70	19 986	802
$\text{bA}_2(3_g)$	$^3\Phi_g$	20 983	180.33	+5.56	20 030	799
$\text{bB}_2(2_g)$	$^3\Delta_g$	21 977	180.12	+5.35	21 121	786
$\text{bA}_1(2_g)$	$^3\Delta_g$	22 537	179.95	+5.18	21 746	778
$\text{cB}_1(3_g)$	$^3\Delta_g$	24 801	179.97	+5.20	23 995	783
$\text{cA}_2(3_g)$	$^3\Delta_g$	24 659	179.96	+5.19	23 862	781
$\text{cB}_2(4_g)$	$^3\Phi_g$	25 665	180.97	+6.20	24 511	789
$\text{cA}_1(4_g)$	$^3\Phi_g$	25 686	180.95	+6.18	24 536	789

^a See footnotes of Tables 3 and 5.

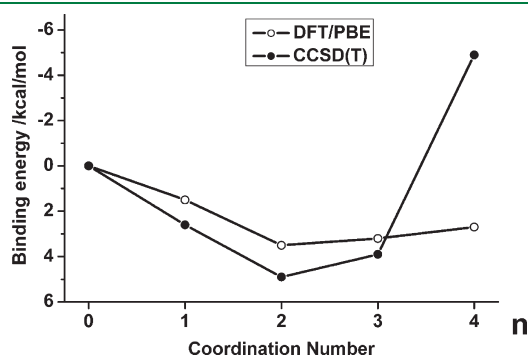


Figure 6. Binding energies of $\text{Ar}_n\text{UO}_2\text{Cl}_2$ ground-state species formed from $\text{UO}_2\text{Cl}_2 + n \text{Ar}$, from CCSD(T) (with BSSE correction) and from DFT/PBE calculations.

Table 7. Geometric Ground-State Parameters of $\text{Ar}_n\text{UO}_2\text{Cl}_2$ Complexes

complex	$R_{\text{U-O}}$, pm	$R_{\text{U-Cl}}$, pm	$\angle \text{OUO}$, °	$\angle \text{ClUCl}$, °	$R_{\text{U-Ar}}$, pm	$R_{\text{U-Ar'}}$, pm
CASPT2						
UO_2Cl_2	175.29	249.12	165.83	106.76		
$\text{Ar}_2\text{UO}_2\text{Cl}_2$	175.47	250.46	168.24	112.19	308.37	
$\text{Ar}_3\text{UO}_2\text{Cl}_2$	175.63	249.53	167.14	98.40	339.92	323.67
CCSD(T)						
UO_2Cl_2	174.77	250.92	165.76	107.73		
$\text{Ar}_2\text{UO}_2\text{Cl}_2$	174.91	252.15	168.17	112.21	312.29	
$\text{Ar}_3\text{UO}_2\text{Cl}_2$	175.03	251.35	166.93	99.34	346.92	327.68

slightly reduced by Ar atoms in the trans position. The Cl–U–Cl angle is reduced by about 8° in $\text{Ar}_3\text{UO}_2\text{Cl}_2$ but unexpectedly widened by some 5° in $\text{Ar}_2\text{UO}_2\text{Cl}_2$.

Since RASSCF/CASPT2-[g_1']/SO and RASSCF/CCSD(T)/SO give consistent results, only the latter are displayed in Table 8 for the cases of zero-, two-, and three-coordinated Ar atoms. The electronic excitation patterns of the three species are quite similar. The argon environment increases the first adiabatic excitation energy by a few 100 cm^{-1} . The higher excitation energies also vary a little. The OUO symmetric stretching frequencies vary in the order of $\pm 10 \text{ cm}^{-1}$, indicating that Ar coordination insignificantly affects these uranyl complexes.

Luminescence of $\text{Ar}_n\text{UO}_2\text{Cl}_2$. The luminescent state properties of $\text{Ar}_n\text{UO}_2\text{Cl}_2$ ($n = 0, 2, 3$) are collected in Table 9. As mentioned before, Ar coordination does not change the dominance of $^3\Phi_g$ character, in contrast to complexes perturbed by carbonyl or carboxyl ligands.^{7,19} Also the U–O bond length changes upon electronic transition, and the vibrational frequencies, i.e. the overall spectral shapes, are insignificantly modified by the Ar ligands. Increases of vertical and adiabatic transition energies of around 450 cm^{-1} for dicoordination and of around 100 cm^{-1} for tricoordination are consistently predicted by RASSCF/CASPT2-[g_1']/SO and RASSCF/CCSD(T)/SO approaches within $\pm 100 \text{ cm}^{-1}$. The electronic transition dipole moments μ of the fluorescence are reduced by factors of 0.4–0.7 due to the Ar coordination.

Heaven et al.⁸ had experimentally found a biexponential decay in the argon matrix with lifetimes of 50 and 260 μs . They had related it to two matrix sites differing in the nonradiative decay times. Assuming the same optical transition probability for both sites, the more rapidly decaying site would have been populated more by a factor of about 4.2. Since the Ar-coordinated complexes have optical transition probabilities differing by factors of $\kappa = 1.6$ –2.1 in the two applied quantum chemical approaches, the deduced population ratio of 4.2:1 should be raised or lowered by that κ factor around 2.

The optical oscillator strengths f are of the order of 10^{-6} , which is common for forbidden transitions in transition-metal complexes. The respective lifetimes are a fraction of a second (s), while the measured ones are a fraction of a ms. Therefore the decay mechanism is dominated indeed by nonradiative transitions. The optical lifetime was estimated by eq 4:¹

$$\tau_{\text{optical}} = 3\pi\epsilon_0\hbar c_n^3 / \omega^3 \mu^2 \quad (4)$$

where ϵ_0 is the electric permittivity in vacuum, \hbar the reduced Planck constant, c_n the velocity of light in solid argon, ω the angular frequency, and μ the electric transition dipole moment.

Table 8. Spectroscopic Data of $\text{Ar}_n\text{UO}_2\text{Cl}_2$ ($n = 0, 2, 3$) from RASSCF/CCSD(T)/SO Calculations^a

state (Ω)	main	UO_2Cl_2		$\text{Ar}_2\text{UO}_2\text{Cl}_2$		$\text{Ar}_3\text{UO}_2\text{Cl}_2$	
		T_e, cm^{-1}	ν_s, cm^{-1}	T_e, cm^{-1}	ν_s, cm^{-1}	T_e, cm^{-1}	ν_s, cm^{-1}
aB ₂ (2 _g)	³ Φ_g	18 797	775	19 364	770	18 942	769
aA ₁ (2 _g)	³ Φ_g	+263	782	+353	776	+332	775
aA ₂ (1 _g)	³ Δ_g	+500	752	+357	752	+432	749
aB ₁ (1 _g)	³ Δ_g	+625	748	+453	756	+498	747
bB ₁ (3 _g)	³ Φ_g	+1189	802	+1220	786	+1192	792
bA ₂ (3 _g)	³ Φ_g	+1233	799	+1184	790	+1266	790
bB ₂ (2 _g)	³ Δ_g	+2324	786	+2308	785	+2284	782
bA ₁ (2 _g)	³ Δ_g	+2949	778	+2670	776	+2837	771
cB ₁ (3 _g)	³ Δ_g	+5198	783	+5106	782	+5106	778
cA ₂ (3 _g)	³ Δ_g	+5065	781	+4910	779	+5066	777
cB ₂ (4 _g)	³ Φ_g	+5714	789	+5887	787	+5818	784
cA ₁ (4 _g)	³ Φ_g	+5739	789	+5917	787	+5823	785

^a See footnotes of Tables 3 and 5.Table 9. Data of the Luminescent States of $\text{Ar}_n\text{UO}_2\text{Cl}_2$ ($n = 0, 2, 3$) from RASSCF/CASPT2-[g_1']/SO and RASSCF/CCSD(T)/SO Calculations^a

molecule	state (Ω)	main	E, cm^{-1}	$\Delta R_e, \text{pm}$	T_e, cm^{-1}	$\mu \times 10^3, \text{au}$	ν_s, cm^{-1}
CASPT2-[g_1']							
UO_2Cl_2	aB ₂ (2 _g)	³ Φ_g	19 471	+5.70	18 509	6.73	752
$\text{Ar}_2\text{UO}_2\text{Cl}_2$	aB ₂ (2 _g)	³ Φ_g	19 844	+5.77	18 869	2.48	747
$\text{Ar}_3\text{UO}_2\text{Cl}_2$	aB ₂ (2 _g)	³ Φ_g	19 539	+5.69	18 591	3.58	744
CCSD(T)							
UO_2Cl_2	aB ₂ (2 _g)	³ Φ_g	19 804	+5.89	18 797	9.99	775
$\text{Ar}_2\text{UO}_2\text{Cl}_2$	aB ₂ (2 _g)	³ Φ_g	20 378	+5.95	19 363	5.94	770
$\text{Ar}_3\text{UO}_2\text{Cl}_2$	aB ₂ (2 _g)	³ Φ_g	19 947	+5.93	18 942	7.44	769

^a See footnotes of Tables 3 and 5, and μ = electronic transition dipole moment between ground and lowest excited states in au at the geometry of the latter.Table 10. Electronic Excitation Spectra of UO_2Cl_2 from Calculations (RASSCF/CASPT2-[g_1']/SO, RASSCF/CCSD(T)/SO) and Experiments^a

state (Ω)	main	T_e, cm^{-1}	ν_s, cm^{-1}	$\mu \times 10, \text{au}$	T_e, cm^{-1}	ν_s, cm^{-1}	$\mu \times 10^3, \text{au}$	T_e, cm^{-1}	ν_s, cm^{-1}	int ^b
		CASPT2-[g_1']			CCSD(T)			exptl. ^c		
XA ₁ (0 _g)			865			901			840	
aB ₂ (2 _g)	³ Φ_g	18 509	752	2.28	18 797	775	5.90	20 359	717	w
aA ₁ (2 _g)	³ Φ_g	+341	760	7.04	+263	782	7.75	+516	706	w
aA ₂ (1 _g)	³ Δ_g	+365	741	0	+500	752	0			
aB ₁ (1 _g)	³ Δ_g	+452	743	0.31	+625	748	0.43			
bB ₁ (3 _g)	³ Φ_g	+1157	772	0.79	+1189	802	0.90	+966	713	w
bA ₂ (3 _g)	³ Φ_g	+1177	776	0	+1233	799	0			
bB ₂ (2 _g)	³ Δ_g	+2197	775	51.42	+2324	786	49.58	+2310	709	s
bA ₁ (2 _g)	³ Δ_g	+2769	766	24.04	+2949	778	22.98	+2522	713	s

^a See footnotes of Tables 3, 5, and 9 (with μ at the ground-state geometry). ^b Int: observed intensity, w = weak, and s = strong. ^c Absorption by UO_2Cl_2 in an Ar matrix.

Assignment and Simulation of Experimental Spectra. Assignment of the Experimental Excitation Spectra. Heaven et al.⁸ had identified 5 progressions in the absorption spectrum of UO_2Cl_2 /argon matrix in the range of $20\text{--}24 \times 10^3 \text{cm}^{-1}$. In Table 10, we

contrast their data with our calculated results for the eight lowest excited states of UO_2Cl_2 in that energy range. Transitions to the two excited states of A₂ symmetry are electronic-dipole forbidden, and the one to the lower state of B₁ symmetry has a distinctively

low calculated intensity; these transitions were too weak to be detected in the mentioned experiment.

Table 11. Vibrational Normal-Mode Coordinates of OOU Symmetric Stretch (*s*) and Bend (*b*) in UO₂Cl₂ from DFT/PBE Frequency Calculations

internal coordinates, unit: Å or Å × rad	normal coordinate, unit: (g/mol) ^{1/2}	
	<i>s</i> (OOU)	<i>b</i> (OOU)
R(U ₁ –O ₂)	2.82	0.04
R(U ₁ –O ₃)	2.82	0.04
R(U ₁ –Cl ₄)	0.17	0.83
R(U ₁ –Cl ₅)	0.17	0.83
∠(O ₂ –U ₁ –O ₃)	0.12	–3.41
∠(O ₂ –U ₁ –Cl ₄)	–0.04	1.01
∠(O ₂ –U ₁ –Cl ₅)	–0.04	1.01
∠(O ₃ –U ₁ –Cl ₄)	–0.04	1.01
∠(O ₃ –U ₁ –Cl ₅)	–0.04	1.01
∠(Cl ₄ –U ₁ –Cl ₅)	0.02	0.88

Table 12. Parameters for Luminescence Spectra Simulation of UO₂Cl₂

parameter ^a	CASPT2-[g ₁ ¹]	CCSD(T)	exptl.
ΔR(U–O), pm	+5.70	+5.89	
Δ(∠OOU), °	–3.33	–5.96	
Δ(∠ClUCl), °	–2.96	–5.57	
Δ(∠OUCl), °	+1.17	+2.13	
ν _s ^g (OOU), cm ^{–1}	865	901	839
ν _s ^e (OOU), cm ^{–1}	752	775	717
ν _b ^g (OOU), cm ^{–1}	185 ^b	214	242
ΔQ _s , Å·(g/mol) ^{1/2}	0.31	0.31	
ΔQ _b , Å·(g/mol) ^{1/2}	0.24	0.42	
T _e , cm ^{–1}	18 509	18 797	20 323
band width, cm ^{–1}			63

^a Changes of geometrical parameters between the ground and luminescent state; ν_s^g and ν_b^g are the symmetric stretching and bending frequencies of the ground state; ν_s^e is the stretching frequency of the luminescent state; and ΔQ_s and ΔQ_b are the corresponding normal coordinate displacements. See footnotes of Table 3 for T_e. ^b From DFT/PBE calculations.

Among the remaining 5 calculated transitions, the lower 3 transitions of (³Φ_g)B₂(2_g), (³Φ_g)A₁(2_g), and (³Φ_g)B₁(3_g) type have medium oscillator strengths of 0.04–3.5 × 10^{–6}, while those of (³Δ_g)B₂(2_g) and (³Δ_g)A₁(2_g) type at higher energy have higher oscillator strengths of 35–167 × 10^{–6}. Indeed, experimentally three weak transitions at lower energies and two strong transitions at higher energies were detected. The calculated energy pattern of these 5 excitations, distributed over nearly 3000 cm^{–1}, is in agreement with experiment within ±300 cm^{–1}, although the calculated absolute excitation energies are on the average ~1400 cm^{–1} too low. This value of less than 0.2 eV is within the accuracy of present-day computational approaches. The calculated excited-state harmonic vibrational frequencies of the symmetric O–U–O stretch are consistently 8% high.

The reasonable agreement corroborates the reliability of the above given assignments of the electronic excitations. With these theoretical results we can improve the previous tentative assignment⁸ with respect to the symmetry species (B₁ or B₂) and the dominant D_{∞h} term values (³Δ_g or ³Φ_g) and angular momenta (Ω = 1, 2, or 3). Namely, the original assignments were based on the calculated vertical excitation energies of UO₂Cl₂-(ac)₃,¹⁹ while we have now found that UO₂Cl₂ in an argon matrix is much less perturbed than in acetone solution.

Simulation of the Experimental Luminescence Spectra. Because the Ar matrix effects on the vibrational frequencies and bond lengths are so weak, we may simulate the shapes of the luminescence spectra with the help of the calculated data for free UO₂Cl₂. Because at present the calculated absolute excitation energies are still not spectroscopically accurate and the prediction of homogeneous and heterogeneous line broadening effects is still too complicated, these two parameters will be adjusted to the experiments.

The two spectroscopically most important, totally symmetric vibrational modes are specified in Table 11. The character is O–U–O stretching and, respectively, a mix of O–U–O and O–U–Cl bending. The other symmetric vibrations, including Cl–U–Cl and Ar–U–Ar stretching and bending modes, are found not to contribute markedly to the shape of the spectra due to the small displacements and/or frequencies.

All numerical parameters for the spectral simulations are assembled in Table 12, and the correspondingly simulated spectra are shown in Figure 7. The experimental emission spectrum⁸ is dominated by a progression of the symmetric O–U–O stretching vibration of 839 cm^{–1} of the electronic ground state. Each band has a weak foot lower by 242 cm^{–1},

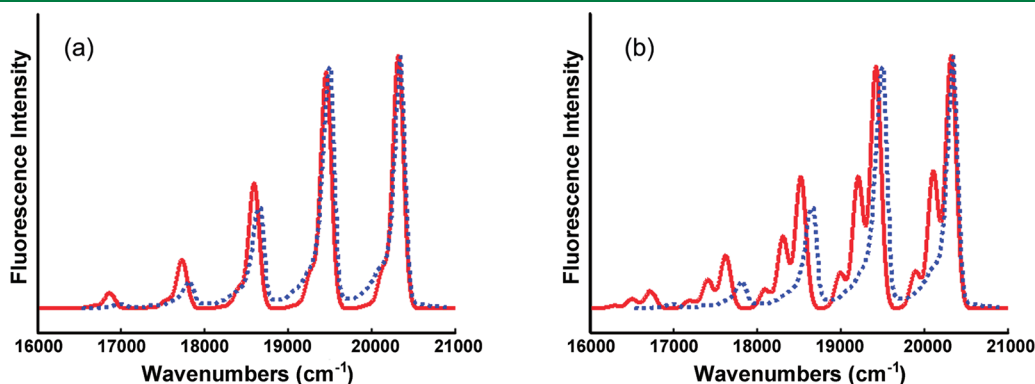


Figure 7. Simulated (red) and experimental (blue) luminescence spectra of UO₂Cl₂: (a): RASSCF/CASPT2-[g₁¹]/SO and (b): RASSCF/CCSD(T)/SO.

corresponding to O–U–O bending. Our calculations corroborate the experimental assignments, with the addition that the bending vibration is a blend of U–O and U–Cl bendings. The CCSD(T) approach seems to exaggerate the intensity of the bending mode, due to the large geometric changes Δ (Table 12), and also yielded the largest ground-state vibrational frequency, which results in the pronounced red shift of the progression (Figure 7).

CONCLUSIONS

We have investigated UO_2Cl_2 in an argon matrix using ab initio RASSCF/CASPT2/SO and RASSCF/CCSD(T)/SO approaches. The electronically excited states and spectra in the near UV–vis region $<27000\text{ cm}^{-1}$ are calculated and interpreted. UO_2Cl_2 interacting with an argon matrix is approximated by $\text{Ar}_n\text{UO}_2\text{Cl}_2$ complexes, where Ar acts as a very weak Lewis base. Two structures ($n = 2$ and 3, Figure 1) are particularly stable, corresponding to the two different matrix sites experimentally deduced and now specified. The luminescence spectra are approximately reproduced theoretically, giving confidence in the theoretical assignments. Accordingly, the fluorescent lowest excited state of UO_2Cl_2 with weakly interacting Ar atoms is of $(\text{U}-5f\phi_u^3\Phi_g)\text{B}_2(2_g)$ character, in contrast to $(\text{U}-5f\delta_u^3\Delta_g)\text{B}_1(1_g)$ of UO_2Cl_2 with equatorial oxygen ligands, such as carbonyl or carboxyl groups. The spectroscopic U–O bond length expansion of UO_2Cl_2 upon electronic excitation from the bonding σ_u -HOMO \rightarrow the nonbonding U-5f ϕ_u -LUMO of $\Delta R_{\text{UO}} \approx 5.8\text{ pm}$ is reproduced by the CASPT2 and CCSD(T) approaches. The bending vibrational frequencies and bond angle changes are theoretically exaggerated by CCSD(T). The absolute intensities, depending both on SO-coupling and geometric asymmetry and on vibronic coupling and radiationless processes, are not calculated. The synergic combination of experimental and theoretical works has helped to eliminate deficiencies of both approaches.

AUTHOR INFORMATION

Corresponding Author

*E-mail: junli@mail.tsinghua.edu.cn.

Present Addresses

[†]Mailing Address of W.H.E.S: Theoretical Chemistry, University of Siegen, 57068 Siegen, Germany. E-mail: schwarz@chemie.uni-siegen.de.

ACKNOWLEDGMENT

The authors thank H. Stoll and M. Dolg for helpful comments. This work was financially supported by the NKBRFSF (2011CB932400) and NSFC (20933003, 11079006, 91026003) of China. The calculations were performed by using computers at the Supercomputing Center of the Computer Network Information Center, Chinese Academy of Sciences, and the Shanghai Supercomputing Center. A portion of calculations was performed using EMSL, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory, Richland, WA.

REFERENCES

(1) Rabinowitch, E.; Belford, R. L. *Spectroscopy and Photochemistry of Uranyl Compounds*; Oxford University Press, Oxford, U.K., 1964.

- (2) Pepper, M.; Bursten, B. E. *Chem. Rev.* **1991**, *91*, 719–741.
- (3) Ghosh, R.; Mondal, J. A.; Ghosh, H. N.; Palit, D. K. *J. Phys. Chem. A* **2010**, *114*, 5263–5270.
- (4) Szabó, Z.; Toraiishi, T.; Vallet, V.; Grenthe, I. *Coord. Chem. Rev.* **2006**, *250*, 784–815.
- (5) (a) Taylor, D. M. *J. Alloys Compd.* **1998**, *271–273*, 6–10. (b) Guilmette, R. A. *J. Alloys Compd.* **1998**, *271–273*, 66–71. (c) Stradling, G. N. *J. Alloys Compd.* **1998**, *271–273*, 72–77. (d) Wolf, S. F. In *The Chemistry of the Actinide and Transactinide Elements*, 3rd ed.; Morss, L. R., Edelstein, N. M., Fuger, J., Katz, J. J., Eds.; Springer: Dordrecht, The Netherlands, 2006; Vol. 5, Chapter 30. (e) Durbin, P. W. In *The Chemistry of the Actinide and Transactinide Elements*, 3rd ed.; Morss, L. R., Edelstein, N. M., Fuger, J., Katz, J. J., Eds.; Springer: Dordrecht, The Netherlands, 2006; Vol 5, Chapter 31.
- (6) (a) Szigethy, G.; Xu, J.; Gorden, A. E. V.; Teat, S. J.; Shuh, D. K.; Raymond, K. N. *Eur. J. Inorg. Chem.* **2008**, 2143–2147. (b) Gorden, A. E. V.; Xu, J.; Szigethy, G.; Oliver, A.; Shuh, D. K.; Raymond, K. N. *J. Am. Chem. Soc.* **2007**, *129*, 6674–6675. (c) Gramer, C. J.; Raymond, K. N. *Inorg. Chem.* **2004**, *43*, 6397–6402. (d) Veeck, A. C.; White, D. J.; Whisenhunt, D. W., Jr.; Xu, J.; Gorden, A. E. V.; Romanovski, V.; Hoffman, D. C.; Raymond, K. N. *Solvent Extr. Ion Exch.* **2004**, *22*, 1037–1068. (e) Xu, J.; Gorden, A. E. V.; Raymond, K. N. *Eur. J. Org. Chem.* **2004**, 3244–3253. (f) Guilmette, R. A.; Hakimi, R.; Durbin, P. W.; Xu, J.; Raymond, K. N. *Radiat. Prot. Dosim.* **2003**, *105*, 527–534. (g) Gorden, A. E. V.; Xu, J.; Raymond, K. N.; Durbin, P. W. *Chem. Rev.* **2003**, *103*, 4207–4282. (h) Xu, J.; Durbin, P. W.; Kullgren, B.; Ebbe, S. N.; Uhlir, L. C.; Raymond, K. N. *J. Med. Chem.* **2002**, *45*, 3963–3971.
- (7) Su, J.; Zhang, K.; Schwarz, W. H. E.; Li, J. *Inorg. Chem.* **2011**, *50*, 2082–2093.
- (8) Jin, J.; Gondalia, R.; Heaven, M. C. *J. Phys. Chem. A* **2009**, *113*, 12724–12728.
- (9) Denning, R. G. *J. Phys. Chem. A* **2007**, *111*, 4125–4143.
- (10) (a) Denning, R. G.; Snellgrove, T. R.; Woodwark, D. R. *Mol. Phys.* **1976**, *32*, 419–442. (b) Barker, T. J.; Denning, R. G.; Thorne, J. R. *G. Chem.* **1987**, *26*, 1721–1732.
- (11) Flint, C. D.; Tanner, P. A. *J. Chem. Soc., Faraday Trans. 2* **1978**, *74*, 2210–2217.
- (12) Görrler-Walrand, C.; de Houwer, S.; Fluyt, L.; Binnemans, K. *Phys. Chem. Chem. Phys.* **2004**, *6*, 3292–3298.
- (13) Li, J.; Bursten, B. E. In *Computational Organometallic Chemistry*; Cundari, T. R., Ed.; Marcel Dekker: New York, 2001; pp 345–379.
- (14) Li, J.; Bursten, B. E. *J. Am. Chem. Soc.* **1998**, *120*, 11456–11466.
- (15) Zhang, Z.; Pitzer, R. M. *J. Phys. Chem. A* **1999**, *103*, 6880–6886.
- (16) Matsika, S.; Pitzer, R. M. *J. Phys. Chem. A* **2001**, *105*, 637–645.
- (17) Pierloot, K.; van Besien, E. *J. Chem. Phys.* **2005**, *123*, 204309.
- (18) Pierloot, K.; van Besien, E.; van Lenthe, E.; Baerends, E. J. *J. Chem. Phys.* **2007**, *126*, 194311.
- (19) van Besien, E.; Pierloot, K.; Görrler-Walrand, C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4311–4319.
- (20) Werner, H. J. MOLPRO, version 2008.1; University College Cardiff Consultants Limited: Cardiff, U.K.; <http://www.molpro.net>
- (21) (a) *Amsterdam Density Functional (ADF) 2009.01*; Scientific Computing and Modeling: Amsterdam, The Netherlands; <http://www.scm.com>. (b) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403. (c) te Velde, G.; Bickelhaupt, F. M.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (22) (a) Weigand, A.; Cao, X.-Y.; Vallet, V.; Flament, J.-P.; Dolg, M. *J. Phys. Chem. A* **2009**, *113*, 11509–11516. (b) Dolg, M.; Cao, X.-Y. *J. Phys. Chem. A* **2009**, *113*, 12573–12581.
- (23) (a) *Pseudopotentials, ECPs*; University of Stuttgart: Stuttgart, Germany. <http://www.theochem.uni-stuttgart.de/pseudopotentials>. (b) Nicklass, A.; Dolg, M.; Stoll, H.; Preuß, H. *J. Chem. Phys.* **1995**, *102*, 8942–8952.
- (24) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (25) (a) van Lenthe, E.; Ehlers, A. E.; Baerends, E. J. *J. Chem. Phys.* **1999**, *110*, 8943–8953. (b) van Lenthe, E.; Baerends, E. J.; Snijders, J. G.

J. Chem. Phys. **1994**, *101*, 9783–9792. (c) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597–4610. (d) van Lenthe, E.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1996**, *105*, 6505–6516. (e) van Lenthe, E.; van Leeuwen, R.; Baerends, E. J.; Snijders, J. G. *Int. J. Quantum Chem.* **1996**, *57*, 281–293.

(26) Wei, F.; Wu, G. S.; Schwarz, W. H. E.; Li, J. *Theor. Chem. Acc.* **2011**, *129*, 467–481.

(27) Malmqvist, P. Å.; Roos, B. O.; Schimmelpfennig, B. *Chem. Phys. Lett.* **2002**, *357*, 230–240.

(28) Roos, B. O.; Malmqvist, P. Å. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2919–2927.

(29) Wei, F.; Wu, G. S.; Schwarz, W. H. E.; Li, J. *J. Chem. Theory Comput.* **2011**, DOI: 10.1021/ct2000233.

(30) Wang, X. B.; Wang, Y.-L.; Yang, J.; Xing, X.-P.; Li, J.; Wang, L. S. *J. Am. Chem. Soc.* **2009**, *131*, 16368–16370.

(31) Wang, Y.-L.; Zhai, H. J.; Xu, L.; Li, J.; Wang, L. S. *J. Phys. Chem. A* **2010**, *114*, 1247–1254.

(32) Wang, Y.-L.; Wang, X. B.; Xing, X.-P.; Wei, F.; Li, J.; Wang, L. S. *J. Phys. Chem. A* **2010**, *114*, 11244–11251.

(33) Fonger, W. H.; Struck, C. W. *J. Chem. Phys.* **1974**, *60*, 1994–2002.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.05, Gaussian Inc.: Pittsburgh, PA, 2003.

(35) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; McGraw-Hill: New York, 1955.

(36) McIntosh, D. F.; Peterson, M. R. General Vibrational Analysis Programs. *Quant. Chem. Prog. Exch. Bull.* **1989**, *9*(3); <http://qcpe@indiana.edu>

(37) McIntosh, D. F. *Theor. Chem. Acc.* **2010**, *125*, 177–184.

(38) Huang, K.; Rhys, A. *Proc. Roy. Soc. London* **1950**, *204A*, 406–423.

(39) Denning, R. G. *Struct. Bonding (Berlin)* **1992**, *79*, 215–276.

(40) (a) Réal, F.; Vallet, V.; Marian, C.; Wahlgren, U. *J. Chem. Phys.* **2007**, *127*, 214302. (b) Réal, F.; Gomes, A. S. P.; Visscher, L.; Vallet, V.; Eliav, E. *J. Phys. Chem. A* **2009**, *113*, 12504–12511.

(41) Wang, Q.; Pitzer, R. M. *J. Phys. Chem. A* **2001**, *105*, 8370–8375.

(42) (a) Li, J.; Bursten, B. E.; Liang, B.; Andrews, L. *Science* **2002**, *295*, 2242–2245. (b) Andrews, L.; Liang, B.; Li, J.; Bursten, B. E. *Angew. Chem., Int. Ed.* **2000**, *39*, 4565–4567.

(43) Li, J.; Bursten, B. E.; Andrews, L.; Marsden, C. J. *J. Am. Chem. Soc.* **2004**, *126*, 3424–3425.

(44) Andrews, L.; Liang, B.; Li, J.; Bursten, B. E. *J. Am. Chem. Soc.* **2003**, *125*, 3126–3139.

(45) Infante, I.; Andrews, L.; Wang, X.; Gagliardi, L. *Chem.—Eur. J.* **2010**, *16*, 12804–12807.

(46) Infante, I.; Kovacs, A.; La Macchia, G.; Shahi, A. R. M.; Gibson, J. K.; Gagliardi, L. *J. Phys. Chem. A* **2010**, *114*, 6007–6015.

(47) (a) Griffiths, D. J. *Introduction to Quantum Mechanics*, 2nd ed; Pearson Education: Upper Saddle River, NJ, 2005; p 356–357. (b) McHale, J. L. *Molecular Spectroscopy*; Prentice Hall: Upper Saddle River, NJ, 1999; p108–109.

Time-Dependent Density Functional Tight Binding: New Formulation and Benchmark of Excited States

Fabio Trani,^{*,†,‡} Giovanni Scalmani,[¶] Guishan Zheng,[§] Ivan Carnimeo,^{†,‡} Michael J. Frisch,[¶] and Vincenzo Barone^{†,‡}

[†]Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126, Pisa, Italy

[‡]INFN Sezione di Pisa, Pisa, Italy

[¶]Gaussian, Inc., 340 Quinipiac Street, Building 40, Wallingford, Connecticut 06492, United States

[§]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

 Supporting Information

ABSTRACT: A new formulation of time-dependent density functional tight binding (TD-DFTB) is reported in this paper. It is derived from the application of the linear response theory to the ground state DFTB Hamiltonian, without the introduction of additional parameters for the description of the excited states. The method is validated for several sets of organic compounds, against the best theoretical estimates from the literature, density functional theory, semiempirical methods, and experimental data. The comparison shows that TD-DFTB gives reliable results both for singlet and triplet excitation energies. In addition, the application of TD-DFTB to open-shell systems shows promising results.

1. INTRODUCTION

Thanks to ongoing developments in hardware, software, and physical models, very reliable results can be obtained for the structure and properties of small- to medium-sized molecules both in their ground state and in excited electronic states. The situation is more involved for large flexible molecules, where some degree of approximation is probably unavoidable. Here, methods rooted in density functional theory (DFT) have revolutionized the situation, paving the way toward a general purpose approach with semiquantitative accuracy along the whole periodic table without the need for prohibitively large basis sets. The subsequent development of the time-dependent (TD-DFT) route to excited electronic states^{1–5} has further enlarged the field of application of quantum-mechanical (QM) approaches. Of course, a number of problems remain to be solved (e.g., true multireference states, charge transfer, van der Waals interactions, etc.),^{6–9} and the development of improved functionals represents a very active research field.^{10–13} The recent literature has shown that, when coupled with suitable functionals (especially hybrid and/or long-range corrected models), the TD-DFT approach provides very good results.^{14–16} The recent availability of TD-DFT analytical gradients both in the gas phase and in solution provides direct access to the structure and properties of excited electronic states at a reasonable computational cost.^{17–19}

However, despite the development of linear scaling methods and other effective techniques, the large systems (with hundreds to thousands of atoms) of biological and technological interest are still highly expensive in terms of computational time. This has stimulated the development, validation, and systematic application of a semiempirical form of density functional theory, i.e., the density functional tight binding method (DFTB),^{20–25} which is several orders of magnitude faster than DFT in practical

calculations. The performances of DFTB for electronic ground states have been well characterized and benchmarked against DFT by several groups.^{26–29} In general, DFTB can well reproduce DFT geometry and energetics, its success being particularly appealing for combined quantum mechanical and molecular mechanics applications.^{30–32} This has stimulated further theoretical developments and more robust parametrizations for wider parts of the periodic table aimed at enlarging the range of application of DFTB.^{33–35}

Motivated by the effectiveness of DFTB and the good performances of TD-DFT, we have developed a new formulation of time-dependent density functional tight binding (TD-DFTB) based on the linear response approach. The scheme is naturally derived from the DFTB equations, and it does not require the introduction of additional parameters, at variance with a previous implementation of the method.³⁶ The present approach has been applied to the calculation of singlet and triplet vertical excitation energies for a benchmark set of organic molecules well studied in the literature.³⁷ The outcomes are then compared with available results of other semiempirical quantum mechanical methods, TD-DFT with different functionals, as well as *ab initio* schemes. In particular, to better understand the difference between DFTB and other semiempirical approaches, TD-PM3 and CIS-PM3 have been performed for the calculations of singlet and triplet excitation energies. The results for a few test molecules are collected and discussed toward experimental data.

In the first section, the DFTB formulation for the ground state is shortly reviewed, followed by the derivation of the TD-DFTB equations. Then, the results for the singlet and triplet excited states of a set of benchmark molecules are reported in

Received: July 4, 2011

Published: September 01, 2011

comparison with the best theoretical estimations from the literature. Statistics for these results are then summarized, and the conclusions are traced, on the basis of the benchmark results.

2. METHOD

In this section, we propose a novel formulation of TD-DFTB based on the linear response approach applied to the DFTB ground state, which is the natural extension of DFTB to the calculation of the excited states. At variance with a previous implementation,³⁶ the present approach (i) does not require the introduction of new or modified parameters and (ii) can be applied to the calculation of excited states of open-shell ground state systems. This makes TD-DFTB a very powerful and computationally affordable method, which can be used in the study of the spectroscopic properties of molecules, nanostructures, and solid crystals. In the following, the derivation of TD-DFTB is presented in detail.

The ground state DFTB total energy is typically written as

$$E = \sum_{i\sigma} \langle \psi_i^\sigma | H^0 | \psi_i^\sigma \rangle + \frac{1}{2} \sum_{AB} \gamma_{AB} \Delta q_A \Delta q_B + \frac{1}{2} \sum_{AII'} p_{AI} W_{AII'} p_{AI'} + \frac{1}{2} \sum_{A \neq B} V_{AB} \quad (1)$$

where ψ_i^σ is the i th occupied molecular orbital (MO) of spin σ , H^0 is the effective core Hamiltonian, and γ_{AB} is a parametrized distance-dependent function²¹ that accounts for the Coulomb interaction between the net atomic charges Δq_A and Δq_B located on atoms A and B . The third term on the right-hand side corresponds to the energy contribution from the atomic spin-densities p_{Ab} , which are partitioned according to the value of the angular momentum l and interact through a one-center atomic parameter $W_{AII'}$. This contribution was not included in the initial formulation of DFTB and was introduced to account for spin densities and to be able to compute hyperfine coupling constants, effectively allowing for unrestricted calculations within DFTB.³⁸ Finally, V_{AB} is a parametrized pairwise interatomic repulsion potential. As other semiempirical methods,^{39–41} each MO is expanded in a valence minimal basis set of atomic orbitals (AO) ($\chi_{i\mu}$)

$$\psi_i^\sigma = \sum_{\mu} c_{i\mu}^\sigma \chi_{i\mu} \quad (2)$$

According to the usual linear combination of atomic orbital (LCAO) scheme, the spin-dependent density matrix and the charge and spin atomic populations are defined as

$$P_{\mu\nu}^\sigma = \sum_i c_{\mu i}^\sigma c_{\nu i}^\sigma \quad (3)$$

$$\Delta q_A = q_A - Z_A = \left[\sum_{\mu \in A} \sum_{\nu} (P_{\mu\nu}^\alpha + P_{\mu\nu}^\beta) S_{\mu\nu} \right] - Z_A \quad (4)$$

$$p_{AI} = \sum_{\mu \in AI} \sum_{\nu} (P_{\mu\nu}^\alpha - P_{\mu\nu}^\beta) S_{\mu\nu} \quad (5)$$

where $S_{\mu\nu}$ are the elements of the overlap matrix and Z_A is the valence nuclear charge of atom A . The ground state DFTB energy in eq 1 is variationally minimized with respect to the LCAO coefficients $c_{\mu i}^\sigma$ using the Kohn–Sham independent particle approximation, and the associate one-particle Fock matrix is defined as

$$F_{\mu\nu}^\sigma = \frac{\partial E}{\partial P_{\mu\nu}^\sigma} = h_{\mu\nu} + G_{\mu\nu}^\sigma \quad (6)$$

where $h_{\mu\nu}$ are the elements of the one-electron operator

$$h_{\mu\nu} = h_{\mu\nu}^0 - \frac{1}{2} S_{\mu\nu} \left[\sum_C (\gamma_{AC} + \gamma_{BC}) Z_C \right] \quad (7)$$

involving the parametrized effective core Hamiltonian matrix elements $h_{\mu\nu}^0$ and the nuclear attraction terms. Both $h_{\mu\nu}^0$ and $S_{\mu\nu}$ are parametrized in DFTB either starting from molecular systems or from bulk reference systems.²¹ The terms in square brackets arise from the interaction between the valence nuclear charge Z_C of atom C and the electron charge density.³⁰

The charge–charge and spin–spin interactions in eq 1 are both quadratic in the density matrix, and they give rise to the two-electron term $G_{\mu\nu}^\sigma$, which can be partitioned into Coulomb- and exchange-like interaction terms

$$G_{\mu\nu}^\sigma = J_{\mu\nu} + K_{\mu\nu}^\sigma \quad (8)$$

where

$$J_{\mu\nu} = \frac{1}{2} S_{\mu\nu} \left[\sum_C (\gamma_{AC} + \gamma_{BC}) q_C \right] \quad (9)$$

$$K_{\mu\nu}^\sigma = \frac{1}{2} \delta_\sigma S_{\mu\nu} \left[\sum_{CII'} (\delta_{AC} W_{CII'} + \delta_{BC} W_{CII'}) p_{CI'} \right] \quad (10)$$

where $\delta_\sigma = \delta_{\sigma\alpha} - \delta_{\sigma\beta}$ (δ_{ij} is the Kronecker delta). The μ and ν AOs are centered on A and B atoms and have l and l' angular momenta, respectively.

In order to apply a linear response approach to the DFTB ground state, it is better to identify the four-index kernels corresponding to the Coulomb and exchange interaction terms,¹⁹ through the second derivatives of the DFTB total energy (eq 1) with respect to the density matrix elements

$$\frac{\partial^2 E}{\partial P_{\mu\nu}^\sigma \partial P_{\kappa\lambda}^\sigma} = J_{\mu\nu\kappa\lambda} + K_{\mu\nu\kappa\lambda}^{\sigma\sigma'} \quad (11)$$

By using a permutation-invariant expression of the total energy derivatives (based on the symmetry properties of the ground state density matrix), the four-index kernels can be written according to the following symmetric form:

$$J_{\mu\nu\kappa\lambda} = \frac{1}{4} S_{\mu\nu} (\gamma_{AC} + \gamma_{BC} + \gamma_{AD} + \gamma_{BD}) S_{\kappa\lambda} \quad (12)$$

$$K_{\mu\nu\kappa\lambda}^{\sigma\sigma'} = \frac{1}{4} \delta_\sigma \delta_{\sigma'} S_{\mu\nu} (W_{AII'} \delta_{AC} + W_{BII'} \delta_{BC} + W_{AII''} \delta_{AD} + W_{BII''} \delta_{BD}) S_{\kappa\lambda} \quad (13)$$

where the μ , ν , κ , and λ AOs are respectively located on A , B , C , and D atoms, and have l , l' , and l'' angular momenta. Since DFTB is derived from a second-order Taylor expansion of the density functional total energy with respect to electron density fluctuations, the DFTB energy can be eventually written as

$$E = \sum_{\sigma} \sum_{\mu\nu} h_{\mu\nu} P_{\mu\nu}^\sigma + \frac{1}{2} \sum_{\sigma\sigma'} \sum_{\mu\nu\kappa\lambda} [J_{\mu\nu\kappa\lambda} + K_{\mu\nu\kappa\lambda}^{\sigma\sigma'}] P_{\mu\nu}^\sigma P_{\kappa\lambda}^{\sigma'} + \frac{1}{2} \sum_{A \neq B} (V_{AB} + Z_A \gamma_{AB} Z_B) \quad (14)$$

The application of the linear response theory to DFTB is straightforward, and it is equivalent to the corresponding derivation within TD-DFT (see, for example, ref 42). The final equations are typically written in superoperator form as a non-Hermitian eigenvalue problem:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \Omega \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \quad (15)$$

where the Ω 's are the excitation energies and the eigenvectors comprise the single excitation $X_{ai} = \delta P_{ai}$ and de-excitation $Y_{ai} = \delta P_{ia}$ amplitudes.⁴² Note that we are using the conventional notation where i, j, \dots are occupied MOs while a, b, \dots are virtual MOs. The elements of the \mathbf{A} and \mathbf{B} matrices are defined as

$$A_{ai\sigma, bj\sigma'} = (\varepsilon_{a\sigma} - \varepsilon_{i\sigma}) \delta_{ab} \delta_{ij} \delta_{\sigma\sigma'} + \frac{\partial F_{ai}^{\sigma}}{\partial P_{bj}^{\sigma'}} \quad (16)$$

$$B_{ai\sigma, bj\sigma'} = \frac{\partial F_{ai}^{\sigma}}{\partial P_{jb}^{\sigma'}} \quad (17)$$

and they involve the ground state orbital energies $\varepsilon_{p\sigma}$. The matrix operator on the left-hand side of eq 15 is typically diagonalized using an iterative scheme, in which the four-index matrices $(\mathbf{A} + \mathbf{B})$ and $(\mathbf{A} - \mathbf{B})$ are evaluated in the AO basis and contracted with the $(\mathbf{X} + \mathbf{Y})$ and $(\mathbf{X} - \mathbf{Y})$ combinations of single excitation amplitudes, once transformed into the AO basis.⁴² Note that in the AO basis, the $(\mathbf{A} + \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ matrix is symmetric with respect to $\mu \leftrightarrow \nu$ and $\kappa \leftrightarrow \lambda$, while the $(\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ is antisymmetric with respect to the same index permutations. Moreover, $(\mathbf{X} + \mathbf{Y})_{\mu\nu\sigma}$ and $(\mathbf{X} - \mathbf{Y})_{\mu\nu\sigma}$ are symmetric and antisymmetric to the $\mu \leftrightarrow \nu$ permutation, respectively. By carrying out the Fock matrix derivatives in eqs 16 and 17 in the AO basis, it is easy to realize that $(\mathbf{A} + \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ involves both charge and the symmetric component of spin interaction kernels:

$$\begin{aligned} (\mathbf{A} + \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'} &\leftarrow (\mathbb{J}_{\mu\nu\kappa\lambda} + \mathbb{J}_{\mu\nu\lambda\kappa}) + (\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} + \mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}) \\ &= 2\mathbb{J}_{\mu\nu\kappa\lambda} + (\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} + \mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}) \end{aligned} \quad (18)$$

while $(\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ involves only the antisymmetric component of the spin interaction kernel:

$$\begin{aligned} (\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'} &\leftarrow (\mathbb{J}_{\mu\nu\kappa\lambda} - \mathbb{J}_{\mu\nu\lambda\kappa}) + (\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} - \mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}) \\ &= (\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} - \mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}) \end{aligned} \quad (19)$$

It is somewhat difficult to understand why the spin interaction kernel has a nonzero antisymmetric component, and therefore an example system only involving the four AO basis functions μ_{Ab} , $\nu_{B'}$, $\kappa_{A'}$, and $\lambda_{B''}$ is given below (subscripts are used to indicate the atom where they are located and their angular momentum label). The two spin interaction kernels corresponding to the $\kappa \leftrightarrow \lambda$ permutation are derived from the general case in eq 13 as

$$\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} = \frac{1}{4} \delta_{\sigma} \delta_{\sigma'} S_{\mu\nu} (W_{A'II''} + W_{B''I''}) S_{\kappa\lambda} \quad (20)$$

$$\mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'} = \frac{1}{4} \delta_{\sigma} \delta_{\sigma'} S_{\mu\nu} (W_{A'II''} + W_{B''I''}) S_{\lambda\kappa} \quad (21)$$

and they are graphically depicted in Figure 1. The expressions in eqs 20 and 21 correspond to the following contribution to the

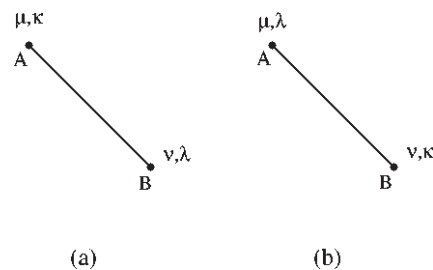


Figure 1. Exchange contributions from the $\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'}$ and $\mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}$ terms, respectively, when the atomic orbitals are centered on atoms A and B.

$(\mathbf{A} + \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ and $(\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ matrices:

$$\begin{aligned} (\mathbf{A} + \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'} &\leftarrow (\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} + \mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}) \\ &= \frac{1}{4} \delta_{\sigma} \delta_{\sigma'} S_{\mu\nu} (W_{A'II''} + W_{A'II''} + W_{B''I''} + W_{B''I''}) S_{\kappa\lambda} \end{aligned} \quad (22)$$

$$\begin{aligned} (\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'} &\leftarrow (\mathbb{K}_{\mu\nu\kappa\lambda}^{\sigma\sigma'} - \mathbb{K}_{\mu\nu\lambda\kappa}^{\sigma\sigma'}) \\ &= \frac{1}{4} \delta_{\sigma} \delta_{\sigma'} S_{\mu\nu} (W_{A'II''} - W_{A'II''} - W_{B''I''} + W_{B''I''}) S_{\kappa\lambda} \end{aligned} \quad (23)$$

where we used the fact that $S_{\kappa\lambda}$ itself is indeed symmetric with respect to the $\kappa \leftrightarrow \lambda$ permutation.

In principle, DFTB is derived from a DFT formalism with a pure exchange-correlation functional, which depends only locally on the density⁴³ (and other density-related quantities like the density gradient etc). In other words, DFTB is derived without the inclusion of nonlocal Hartree–Fock exact exchange, and therefore, just as in the case of TD-DFT using pure functionals, the $(\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ matrix should be strictly diagonal.⁴² The previously published formulation of TD-DFTB³⁶ is consistent with this observation because the on-site spin–spin interaction parameters W_A used in that paper are scalars, which do not depend on the AO angular momentum. Under these conditions, $(\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ does remain strictly diagonal since eqs 20 and 21 become identical, and the right-hand side of eq 23 vanishes. However, the parameters being used for the spin–spin interaction in an, e.g., a triplet excited state computed as an excitation from a closed-shell singlet are not the same that would have been used if that triplet was computed using the open-shell formulation of DFTB for the ground state³⁸ (where the $W_{A'II''}$ are not scalars, as they depend on the AO angular momentum).

To summarize, in this section, we have shown how to apply the linear response approach to an unrestricted ground state DFTB calculation. The resulting TD-DFTB method does not require additional or modified spin–spin parameters, and it is applicable to both closed- and open-shell ground states. Notably, in our formulation of TD-DFTB, the $(\mathbf{A} - \mathbf{B})_{\mu\nu\sigma\kappa\lambda\sigma'}$ matrix is not diagonal, just like what happens in the case of TD-DFT when an hybrid functional is used, i.e., when a fraction of the nonlocal Hartree–Fock exact exchange is present. Indeed, this suggests that the open-shell formulation of DFTB which involves orbital angular momentum dependent spin–spin parameters $W_{A'II''}$ could be already regarded as a model derived from a hybrid DFT formalism which involves the exact exchange.

Table 1. Sets of Organic Compounds Used for Benchmark

group I	group II	group III	group IV
ethene	benzene	formaldehyde	cytosine
E-butadiene	naphthalene	acetone	thymine
all-E-hexatriene	furan	p-benzoquinone	uracil
all-E-octatetraene	pyrrole	formamide	adenine
cyclopropene	imidazole	acetamide	
cyclopentadiene	pyridine	propanamide	
norbornadiene	pyrazine		
	pyrimidine		
	pyridazine		
	s-triazine		
	s-tetrazine		

3. RESULTS

The method described in the previous section has been used to calculate the excitation energies in the benchmark set of organic compounds proposed by Thiel and co-workers.³⁷ In that paper, the authors collected the Theoretical Best Estimates (TBEs) for both singlet and triplet vertical excitation energies calculated at the CASPT2 level with a molecular structure optimized at the MP2 level. The same sets of compounds and excitations have been used to benchmark the accuracy of multireference Configuration Interaction (CI) approaches,¹⁵ TD-DFT (by assessing the performance of many exchange-correlation functionals),^{16,44,45} and semiempirical quantum chemical methods.⁴⁶ For this reason, the same four sets of compounds and excitations are used in this paper as a benchmark for the TD-DFTB scheme. The compounds are divided into four groups, as summarized in Table 1. Group I contains unsaturated aliphatic hydrocarbons; group II, aromatic hydrocarbons and heterocycles; group III, aldehydes, ketones, and amides; and group IV, nucleobases.

The optimized geometries at the MP2 level were taken from the literature.³⁷ Indeed, the use of good quality MP2-optimized geometries seems to be accepted as a standard approach to benchmark the accuracy of a method for excited state calculations, since it avoids biasing the results due to the quality of the ground-state geometries provided by the method under testing. But, in order to show the good quality of the method, TD-DFTB calculations using geometries optimized at the DFTB electronic ground state were also performed. In the Supporting Information, the full set of singlet and triplet vertical excitation energies are collected, using geometries optimized both at the MP2 level and at the DFTB level.

TD-DFTB is based on a minimal orbital basis set that can only model valence excited states, while it cannot describe diffuse excitations such as Rydberg states. For this reason, the analysis was restricted to valence excited states. To better compare the method against standard semiempirical schemes, TD-PM3 and CIS-PM3 calculations on singlet and triplet excited states have been performed for a restricted set of molecules. This set has been recently used to carefully validate TD-DFT toward experimental data.⁴⁷ A comparative study of the methods (TD-PM3, TD-DFTB, TD-DFT) against TBEs and experimental data is discussed for this restricted set of molecules, before going to the statistics of TD-DFTB on the large benchmark set of molecules listed in Table 1. All calculations were performed using

TD-DFTB, TD-PM3, and CIS-PM3 implemented in the development version of the Gaussian program.⁴⁸

Singlet States. Singlet excitation energies have been computed using TD-DFTB with geometries optimized at the MP2 level. Before reporting on the statistics obtained from the benchmark set of molecules, it is interesting to compare the excitation energies obtained by TD-DFTB with the ones obtained by other semiempirical methods. The comparison is done using as a reference a recent paper where configuration interaction calculations using single, double, triple, and quadruple excitations (CISDTQ) have been proposed for standard semiempirical methods (MNDO, INDO/S, AM1, PM3) and orthogonalization-corrected approaches (OM1, OM2, OM3).⁴⁶ To better compare TD-DFTB and semiempirical approaches, we applied TD-PM3 and CIS-PM3 to calculate singlet excitation energies for a few molecules recently carefully characterized using DFT approaches.⁴⁷ The results have been collected in Table 2, where TD-DFTB is compared to TD-PM3, CIS-PM3, experimental data, TD-PBE, TD-B3LYP, TBEs, and CISDTQ-PM3.

As can be observed from the table, the singlet excitation energies calculated at the PM3 level are often strongly underestimated within TD and CIS approximation. As a matter of fact, semiempirical approaches often need a CISDTQ level of approximation for obtaining converged excitation energies, thus requiring a high degree of computational effort. Table 2 shows that, with respect to PM3, in most cases, TD-DFTB is much closer to TBE/experimental data. For instance, the excitation energy of ethene ${}^1B_{1u}$ is strongly underestimated within TD-PM3 (it is more than 2.0 eV below). The energy is raised when using CIS-PM3 (5.86 eV) and then CISDTQ-PM3 (6.63 eV), but even in this case, the error with respect to TBE (7.80 eV) and experimental data (7.65) is greater than 1 eV. On the other hand, the excitation energy is well reproduced by TD-DFTB (7.64 eV).

The trend TD-PM3 < CIS-PM3 < CISDTQ-PM3 < TBE/exptl is observed for ethene, E-butadiene, formaldehyde, and acetone. The situation changes for aromatic hydrocarbons, where, however, apart from a few cases, TD-DFTB performs much better than PM3. For instance, the pyrazine ${}^1B_{3u}$ excited state is evaluated at 3.31, 3.35 and 3.29 eV within TD-, CIS- and CISDTQ-PM3, respectively, against 3.95 eV (TBE) and 3.83 eV (experimental). At the same MP2-level optimized geometry, TD-DFTB gives a better estimate at 3.66 eV.

Standard semiempirical methods of quantum chemistry give results even worse when using PM6 instead of PM3. In particular, TD-PM6 gives 4.98 and 3.24 eV for ethene ${}^1B_{1u}$ and pyrazine ${}^1B_{3u}$ excited states, respectively, thus showing a significant underestimation of singlet excited states with respect to all other approaches here considered. From these comparisons, TD-DFTB emerges as an invaluable tool for predicting excited states of organic molecules. In the last rows of Table 2, the mean errors and standard deviations with respect the experimental data are reported. It is worth noting that these data are merely indicative since the set of considered compounds is really small. It only involves the lowest excitations, and it is unbalanced toward the aromatic hydrocarbons. For this reason, TD-DFTB appears with a small value of the signed error with respect to the experimental data. Nonetheless, Table 2 gives important information to compare in a glance the several methods for a few molecules, with respect to the experimental data. From now on, this section will focus on the larger benchmark set, composed by the compounds listed in Table 1, and more suitable for doing statistics. All of the excitation energies for the benchmark set,

Table 2. Comparison of Singlet Vertical Excitation Energies (in eV) Calculated for a Few Molecules Using Different Approaches^a

		PM3			DFTB	PBE	B3LYP		exptl
		TD	CIS	CISDTQ	TD	TD	TD	TBE	
ethene	¹ B _{1u} (π → π*)	5.68	5.86	6.63	7.64	7.07	7.32	7.80	7.65
E-butadiene	¹ B _u (π → π*)	4.60	4.78	5.45	5.49	5.41	5.54	6.18	5.91
formaldehyde	¹ A ₂ (n → π*)	2.65	2.69	2.87	4.13	3.78	3.89	3.88	4.00
	¹ B ₁ (σ → π*)	8.47	8.48	8.63	8.18	8.83	8.96	9.10	9.00
acetone	¹ A ₂ (n → π*)	3.08	3.12	3.29	4.38	4.20	4.36	4.40	4.43
pyridine	¹ B ₁ (n → π*)	3.75	3.79	3.75	4.49	4.34	4.78	4.59	4.59
	¹ B ₂ (π → π*)	4.04	4.17	3.35	5.31	5.31	5.44	4.85	4.99
	¹ A ₂ (n → π*)	4.32	4.33	3.96	4.84	4.44	5.10	5.11	5.43
pyrazine	¹ A ₁ (π → π*)	4.28	4.34	5.00	5.77	6.17	6.23	6.26	6.38
	¹ B _{3u} (n → π*)	3.31	3.35	3.29	3.66	3.55	3.94	3.95	3.83
	¹ B _{2u} (π → π*)	4.12	4.26	3.48	5.11	5.22	5.31	4.64	4.81
	¹ B _{2g} (n → π*)	4.09	4.13	4.13	5.39	5.11	5.56	5.56	5.46
pyrimidine	¹ B _{1g} (n → π*)	5.19	5.19	4.51	6.14	5.57	6.40	6.60	6.10
	¹ B ₁ (n → π*)	3.53	3.57	3.46	4.17	3.77	4.26	4.55	3.85
	¹ A ₂ (n → π*)	3.84	3.86	3.66	4.52	3.99	4.59	4.91	4.62
pyridazine	¹ B ₂ (π → π*)	4.24	4.36	3.55	5.48	5.57	5.71	5.44	5.12
	¹ B ₁ (n → π*)	3.14	3.18	3.15	3.51	3.11	3.56	3.78	3.60
	¹ A ₂ (n → π*)	4.36	4.40	4.16	4.94	5.02	5.46	5.77	5.30
s-tetrazine	¹ A ₁ (π → π*)	4.11	4.24	3.41	5.30	5.44	5.58	5.18	5.00
	¹ B _{3u} (n → π*)	2.32	2.36	2.27	2.39	1.83	2.24	2.29	2.25
	¹ A _u (n → π*)	3.26	3.30	2.99	3.56	2.80	3.48	3.51	3.40
mean error		−0.92	−0.86	−0.99	−0.06	−0.25	0.09	0.13	
std. dev.		0.54	0.52	0.50	0.34	0.38	0.29	0.25	

^a All calculations have been performed at MP2-level optimized geometries, as given in ref 37. Experimental data and TD-DFT results have been taken from Ref 47, while CISDTQ results have been taken from ref 46. At the bottom of the table, the signed mean error and the standard deviation with respect to the experimental data are reported. The last column contains the statistics on the whole set.

Table 3. Statistics for Vertical Singlet Excitation Energies (eV) Calculated within TD-DFTB for the Four Groups of Compounds, with Respect to the TBES

	group I	group II	group III	group IV	total
count	13	53	19	19	104
mean	−0.46	−0.28	−0.31	−0.90	−0.42
abs. mean	0.46	0.34	0.54	0.90	0.49
std. dev.	0.30	0.34	0.57	0.52	0.48
max + dev.	−0.05	0.47	0.61	−0.27	0.61
max − dev.	0.90	1.05	1.19	1.94	1.94

calculated within TD-DFTB, are reported in the Supporting Information.

Statistics on the four groups of compounds with respect to the TBES are reported in Table 3. The best agreement between TD-DFTB and the TBES is found for group II, i.e., for the aromatic hydrocarbons and heterocycles, while the worst agreement is undoubtedly found for the nucleobases (group IV). This is in line with the results obtained using other semiempirical quantum chemical methods, such as AM1 and PM3.⁴⁶

The results obtained using geometries optimized at both the MP2 and DFTB levels have been compared (see the Supporting Information). While excitation energies can be sometimes slightly different for the two cases, we found that overall the statistics do not change much. The signed and absolute mean

errors for excitation energies obtained using the DFTB-optimized geometries are −0.40 and 0.50 eV, in agreement with the results obtained with the MP2-optimized geometries, that are −0.42 and 0.49 eV, respectively (see Table 3).

A comparison with the results provided by a previous implementation³⁶ of the TD-DFTB method has been carried out, and a good agreement is found. The differences between the current and the previous approach are quite small, usually below 0.1 eV (see Supporting Information). As shown above, the main difference between the two schemes is in the antisymmetric contribution through the (A − B) matrices, due to the angular dependence of the spin–spin interaction parameters. The antisymmetric contribution is quite small for singlet excited states, and this justifies the observed agreement between the two different formulations of TD-DFTB.

The present results (Table 3) show that TD-DFTB tends to underestimate the excitation energies. The signed mean error (the deviation from TBE) for singlet excited states is −0.42 eV, which is consistent with the fact that the parametrization of the DFTB Fock matrix was done on the basis of DFT calculations using the PBE functional, which underestimates (on average) the singlet excitation energies of organic molecules.^{16,47}

On the other hand, TD-DFTB performs much better than CISDTQ-PM3 and -AM1, whose mean errors are −1.40 eV and −1.14 eV, respectively.⁴⁶ The present method shows a mean error larger than INDO/S2 (−0.11 eV), but in those calculations the errors appear to be spread (the standard deviation is 0.77 eV).

TD-DFTB results are less scattered with respect to the INDO results, and the standard deviation of 0.48 eV is even smaller than the one observed for OMx methods (OM3 shows a standard deviation of 0.54 eV).⁴⁶

The deviations for singlet excitation energies are shown in Figure 2. With the exception of a few cases, a systematic underestimation of the excitation energies is apparent, which can be due to the behavior of the PBE functional, which is inherited by DFTB through its parametrization. With few exceptions, TD-DFTB gives errors which are all scattered within a fairly small energy range of about 2 eV.

Triplet States. The TD-DFTB method introduced in the previous section has been benchmarked against TBEs for singlet–triplet vertical excitation energies. All of the results have been collected in the Supporting Information. In Table 4, a comparison of the method toward TD-, CIS-, and CISDTQ-PM3 for a few selected molecules is reported.

The trend TD-PM3 < CIS-PM3 < CISDTQ-PM3 < TBE/exptl, observed in the previous section, is confirmed for triplet excitation energies of ethene, E-butadiene, formaldehyde, and

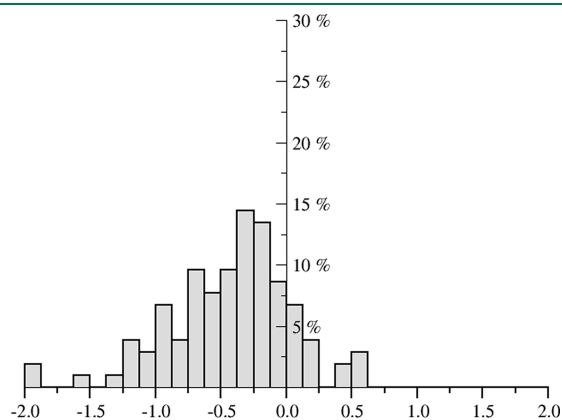


Figure 2. Histograms of the deviations for singlet vertical excitation energies (in eV) calculated by TD-DFTB with respect to the TBEs.

acetone. For instance, the ethene ${}^3B_{1u}$ excitation energy is 1.54, 2.56, and 3.05 eV, within TD-, CIS-, and CISDTQ-PM3 approximations, respectively, against the reference value of 4.50 eV (TBE). Therefore, in the best case, PM3 is about 1.5 eV below the reference value. Instead, the excitation energy reported within TD-DFTB is 5.26 eV, that is, 0.76 eV above TBE. A similar situation is observed for formaldehyde, where the best PM3 estimate of 3A_2 excitation energy, 2.57 eV, obtained within CISDTQ, is 0.93 eV below TBE/exptl (3.50 eV). TD-DFTB instead gives 4.13 eV, that is, 0.63 eV above the reference value. The same trend is observed for pyridine, while for s-tetrazine, PM3 gives better results. In fact, while for the other molecules here considered, PM3 (within all approximations) gives results well below both TBE and experimental data; for s-tetrazine, there is better agreement. Just like for singlet excitations, even for triplet excited states, a careful analysis has been performed on a large benchmark set, with the aim of getting realistic statistical data. A table containing all of the triplet excitation energies is included in the Supporting Information file.

In Table 5, the statistics on the four groups of organic compounds reported in Table 1 are summarized. A histogram with the deviations for the excitations toward triplet states is shown in Figure 3. An interesting feature of TD-DFTB emerges from this analysis. Other semiempirical approaches either show the same precision (PM3, OMx) or they perform much worse (INDO) for triplet states than for singlet states.⁴⁶ At variance with them, TD-DFTB gives better results (on average) for triplet than for singlet excited states. Indeed, the signed mean error is very small (0.07 eV). The absolute error is also smaller with respect to the singlet excitation energies, and the deviations from the TBE values are less scattered (the standard deviation is 0.43 eV). Our results indicate that while TD-DFTB quite systematically underestimates the excitation energies of singlet states, the signed deviations from the TBEs for triplet states sum up to a very small mean error.

Another interesting point of comparison is the first singlet to triplet energy difference $S_0 \rightarrow T_0$. The energy of the T_0 state can be computed in two ways: either using TD-DFTB and therefore

Table 4. Singlet–Triplet Vertical Excitation Energies (in eV) Calculated for a Few Molecules Using Different Approaches^a

		PM3			DFTB	PBE	B3LYP	TBE	exptl
		TD	CIS	CISDTQ	TD	TD	TD		
ethene	${}^3B_{1u}(\pi \rightarrow \pi^*)$	1.54	2.56	3.05	5.26	4.25	4.05	4.50	4.6
E-butadiene	${}^3B_u(\pi \rightarrow \pi^*)$		1.87	2.28	3.62	2.95	2.78	3.20	3.22
	${}^3A_g(\pi \rightarrow \pi^*)$	2.22	2.86	3.32	5.53	5.03	4.87	5.08	4.91
formaldehyde	${}^3A_2(\pi \rightarrow \pi^*)$	2.27	2.31	2.57	4.13	3.02	3.13	3.50	3.50
	${}^3A_1(\pi \rightarrow \pi^*)$	4.28	4.73	5.07	6.50	5.56	5.20	5.87	5.82
acetone	${}^3A_2(n \rightarrow \pi^*)$	2.72	2.76	3.05	4.38	3.55	3.69	4.05	4.16
	${}^3A_1(\pi \rightarrow \pi^*)$	3.92	4.35	4.65	6.19	5.63	5.40	6.03	5.88
pyridine	${}^3A_1(\pi \rightarrow \pi^*)$		2.42	2.68	4.77	4.14	3.90	4.06	4.10
	${}^3B_2(\pi \rightarrow \pi^*)$	3.46	3.51	3.44	4.81	4.45	4.52	4.64	4.84
s-tetrazine	${}^3B_{3u}(n \rightarrow \pi^*)$	1.77	1.84	1.88	2.39	1.09	1.42	1.89	1.69
	${}^3A_u(n \rightarrow \pi^*)$	2.86	2.92	2.80	3.56	2.48	3.10	3.52	2.90
	${}^3B_{1g}(n \rightarrow \pi^*)$	2.96	3.00	3.14	4.63	3.30	3.63	4.21	3.60
mean error		−1.39	−1.17	−0.94	0.55	−0.31	−0.29	0.11	
std. dev.		1.02	0.71	0.57	0.28	0.22	0.25	0.26	

^a All calculations have been performed at MP2-level optimized geometries, as given in ref 37. Experimental data and TBEs have been taken from ref 37. TD-PBE and TD-B3LYP calculations are from ref 44. CISDTQ results are from ref 46. At the bottom of the table, the signed mean error and the standard deviation with respect to the experimental data are reported. All values are in eV.

Table 5. Deviations in the Vertical Excitation Energies (eV) Towards Triplet States within TD-DFTB For All the Compounds in the Benchmark Set, as Compared with the TBEs

	group I	group II	group III	total
count	13	36	14	63
mean	0.37	-0.02	0.03	0.07
abs. mean	0.39	0.36	0.30	0.35
std. dev.	0.27	0.44	0.41	0.43
max + dev.	0.77	0.71	0.63	0.77
max - dev.	0.18	1.08	0.87	1.08

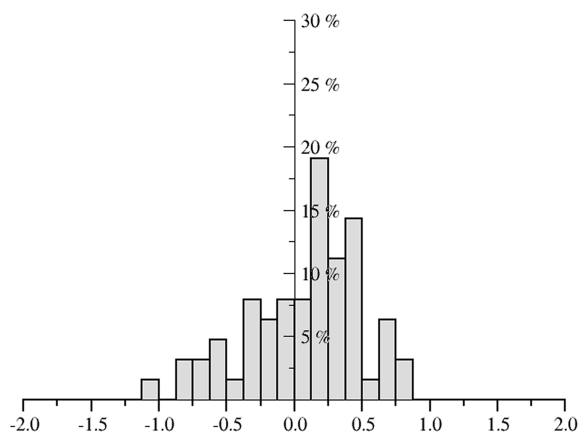


Figure 3. Histograms of the deviations for triplet vertical excitation energies (in eV) calculated by TD-DFTB with respect to the TBEs.

computing T_0 as an excited state of S_0 or by performing an unrestricted ground state DFTB calculation of T_0 . When using a valence minimal basis set, the Hartree–Fock method and other semiempirical methods like PM6 provide the same energy for the T_0 irrespective of the way the latter is computed. This is not the case for DFTB, because all of the on-site exchange integrals between functions with different angular momenta vanish due to the Neglect of Differential Overlap (NDO) approximation, which makes each atomic sub-block of $S_{\mu\nu}$ diagonal. This appears to be a limitation in the current formulation of DFTB which effectively lifts the degeneracy of the three components of a triplet state.

All calculations of triplet excitation energies were performed using geometries optimized at the MP2 level, and the results have been collected in the Supporting Information. The analysis of the results does not allow for the identification of any general trend. For example, in the case of ethene, the excitation energy toward the ${}^3B_{1u}$ triplet state is predicted at 5.26 eV by TD-DFTB, while a triplet ground state calculation leads to a energy of 5.44 eV. These two results are to be compared with a TBE reference energy of 4.50 eV. In the case of formaldehyde, the singlet to triplet transition energy is computed at 4.13 eV by TD-DFTB, while the triplet ground state is at 3.71 eV and the reference value is 3.50 eV.

As in the case of the singlet excitation energies, the structural optimization by means of DFTB has a small effect, on average. The comparison with excitation energies obtained using a DFTB-optimized structures leads to 0.14 and 0.39 eV for signed and absolute error, respectively. This is in agreement with the results

obtained using the geometries optimized at the MP2 level. Also, for triplet excitation energies, the results were compared with the ones published in the previous implementation of the method,³⁶ and fair agreement (see Supporting Information) is found. Finally, it is observed that neglecting the antisymmetric contribution from $(A - B)$ has only a minor effect on the triplet excitation energies, corresponding to a change typically within 0.02 eV. Just like for singlet excited states, TD-DFTB gives good results for aromatic hydrocarbons (group II), with an overall compensation of the mean error. On the other hand, the method tends to overestimate the energies for group I molecules, but in this case, the standard deviation is small and the results are less scattered.

4. STATISTICS

In Figure 4 are shown the correlation plots of the TD-DFTB results against the TBEs, for both singlet (left panel) and triplet (right panel) excited states, for the large benchmark set of Table 1. The plots summarize the findings discussed in the previous section. Singlet vertical excitation energies are systematically underestimated within the whole range of energies considered here. The best-fit curve obtained by linear regression is parallel to the bisector, lying about 0.4 eV below it. For triplet states, the slope is slightly different, but in the energy range considered, the difference does not appear to be significant. The correlation coefficient between TD-DFTB excitation energies and TBE is $r = 0.944$ and $r = 0.940$ for singlet and triplet states, respectively. The correlation plots and the corresponding correlation coefficients are either superior or comparable with respect to the ones obtained with other semiempirical methods, such as PM3 ($r = 0.868$ and $r = 0.880$ for singlet and triplet states, respectively), INDO/S2 ($r = 0.897$ and $r = 0.796$), and OM3 ($r = 0.932$ and $r = 0.925$).⁴⁶ The high correlation coefficient shows that, in practice, the TD-DFTB singlet excitation energies do have a quality comparable to the best theoretical estimates from the literature when a rigid shift of 0.4 eV is applied.

A graphical comparison with other methods is shown in Figure 5, where the mean deviation of different methods with respect to TBEs is shown, with an error bar corresponding to the value of the standard deviation. Statistics for semiempirical methods have been taken from Thiel and Silva-Junior.⁴⁶ All of the results in black-closed symbols come from statistics performed with respect to the theoretical best estimates from the literature, over the large benchmark set of 104 singlet and 63 triplet excited states, of the organic molecules grouped in Table 1 and first proposed in ref 37. In addition, the statistics of Table 2, which refer to TD-DFTB and TD-DFT energies with respect to the experimental data, have been added to the figure as red-open symbols. They refer to a small benchmark set, but the convergence of TD-DFT results has been carefully checked by using a large basis set.⁴⁷ It is therefore included for completeness (M05 results have also been added).

As far as singlet excited states (left panel) are concerned, TD-DFTB is in fair agreement with the PBE functional. This is likely due to the fact that the DFTB parametrization of the Fock matrix has been carried out on referenced PBE calculations. TD-DFTB can undoubtedly be preferred over INDO, where significant error compensation leads to a small deviation from TBE, but the spread (the error bar in the graph) is large. Moreover, TD-DFTB appears to be more accurate than PM3, which strongly underestimates the excitation energies. Furthermore, TD-DFTB

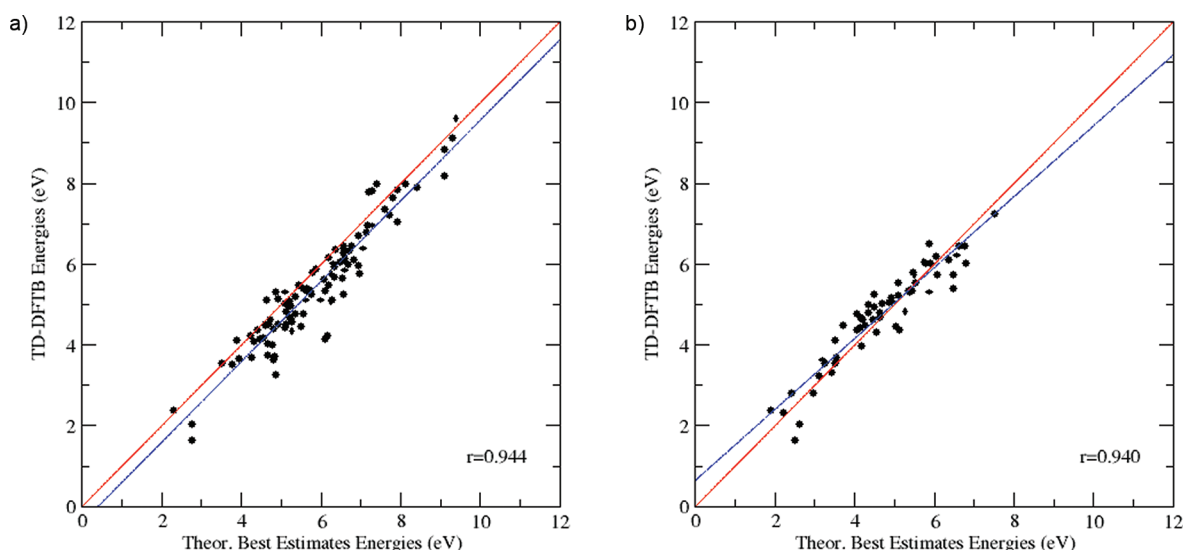


Figure 4. Correlation plots of TD-DFTB vertical excitation energies for singlet states (left panel) and triplet states (right panel) with respect to the TBEs.

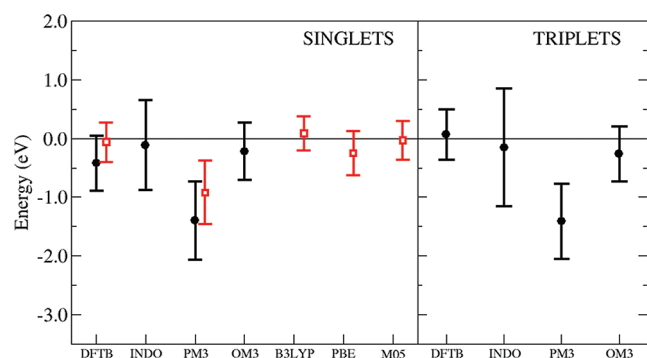


Figure 5. Graphical comparison with other methods. Black-closed symbols: mean errors of the excitation energies with respect to the TBEs, calculated for the whole benchmark set of molecules in Table 1. TD-DFTB results are compared to other results obtained from the literature, CISDTQ-PM3, INDO/S2, and OM3,⁴⁶ for singlet and triplet excited states. Red-open symbols: mean errors with respect to the experimental data of TD-DFTB, TD-PM3, TD-B3LYP, and TD-PBE data (see Table 2), calculated on a small benchmark set of excitations. TD-MO5 results on the same set have been taken from ref 47.

provides results of comparable accuracy to the ones obtained by the semiempirical OM3 approach, but at a much lighter computational effort.

The comparison of the results obtained for triplet excited states is shown in the right panel of Figure 5. In the case of standard semiempirical methods, the statistics are very similar to the ones obtained for singlet states. Instead, from this figure, it is interesting to note that TD-DFTB performs better for triplet than for singlet excited states (compare TD-DFTB black-symbols, left and right panels). It has a smaller spread and a smaller signed error from TBE. It is important to remark that this is overall statistics over a large set of molecules and excited states. As was observed in the previous section (see also Supporting Information), DFTB tends to overestimate the low-lying triplet excitation energies (see Table 4), this trend being compensated from an underestimation of high-lying excitations. More than the

Table 6. Vinyl Radical Excited States (eV) Calculated by TD-B3LYP⁴⁹ and by TD-DFTB, Using the B3LYP-Optimized Geometry in Both Cases

excited state	transition	TD-B3LYP	TD-DFTB
$^2A''$	$\pi \rightarrow n$	3.29	2.84
$^2A''$	$n \rightarrow \pi^*$	4.48	4.22
$^2A'$	$\pi \rightarrow \pi^*$	4.40	4.93

mean signed error, the standard deviation (error bar in figure) is another important indicator in qualifying the goodness of a method, and TD-DFTB has a standard deviation that is much smaller than INDO and PM3 and comparable to OM3. To conclude, Figure 5 actually shows that DFTB is the best semiempirical approach, among the different models here considered, for the calculation of triplet excitation energies of organic molecules.

As a final benchmark, the excited states of the vinyl radical were calculated, as a simple case of a ground-state open-shell system, to check how TD-DFTB performs in doublet excited states calculations. A recent paper reports accurate DFT calculations of the excited states of the vinyl radical; thus we have chosen B3LYP results as a reference for TD-DFTB.⁴⁹ Among the low-lying excitations, two excited states correspond to the $\pi \rightarrow n$ and $n \rightarrow \pi^*$ transitions, while a third excited state corresponds to a $\pi \rightarrow \pi^*$ transition (note that the vinyl radical has an n -type single-occupied molecular orbital). TD-DFTB calculations were performed using the geometry optimized at the B3LYP level, and the vertical excitation energies for the first doublet states are reported in Table 6. We found an underestimation of the first ($\pi \rightarrow n$) and the second ($n \rightarrow \pi^*$) excited states by 0.45 eV and 0.26 eV, respectively. On the other hand, the A' state is overestimated by 0.53 eV using TD-DFTB.

5. CONCLUSIONS

A new formulation and implementation of the time-dependent density functional tight binding method (TD-DFTB) has

been introduced in this paper. The present scheme follows from the linear response theory applied to the DFTB ground state equations. At variance with a previously developed formulation³⁶ of TD-DFTB, the current approach does not require any additional or modified parameters for the calculation of the excited states, and it is characterized by a nontrivial antisymmetric spin contribution in the TD-DFTB equations. The current version of TD-DFTB has been benchmarked against theoretical best estimates and experimental data, from which promising results are found. It has been shown that TD-DFTB provides accurate prediction of transition energies toward both singlet and triplet excited states. To compare the method against the standard semiempirical approaches, TD-PM3 and CIS-PM3 calculations have been performed on a small set of molecules, and the results are compared to CISDTQ-PM3 results from the literature. It has been found that TD-DFTB performs better than other standard semiempirical quantum chemical methods.

■ ASSOCIATED CONTENT

S Supporting Information. All of the vertical transition energies for singlet and triplet excitations, obtained by means of TD-DFTB, using molecular geometries optimized both at the MP2 and at the DTFB level, are available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fabio.trani@sns.it.

■ ACKNOWLEDGMENT

Part of the calculations were performed using the CINECA advanced computing facilities.

■ REFERENCES

- (1) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.
- (2) Casida, M. E. *Recent Advances in Density Functional Methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Vol. 1, pp 155–192.
- (3) Gross, E.; Dobson, J.; Petersilka, M. Density functional theory of time-dependent phenomena. In *Density Functional Theory II*; Nalewajski, R., Ed.; Springer: Berlin/Heidelberg, 1996; Vol. 181, pp 81–172.
- (4) Petersilka, M.; Gossmann, U. J.; Gross, E. K. U. *Phys. Rev. Lett.* **1996**, *76*, 1212–1215.
- (5) Marques, M.; Gross, E. *Annu. Rev. Phys. Chem.* **2004**, *55*, 427–455.
- (6) Burke, K.; Werschnik, J.; Gross, E. K. U. *J. Chem. Phys.* **2005**, *123*, 062206.
- (7) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.
- (8) Casida, M. E. *THEOCHEM* **2009**, *914*, 3–18.
- (9) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007–4016.
- (10) Tozer, D.; Amos, R.; Handy, N.; Roos, B.; Serrano-Andres, L. *Mol. Phys.* **1999**, *97*, 859–868.
- (11) Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151–160.
- (12) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (13) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.
- (14) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *302*, 375–382.
- (15) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.
- (16) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.
- (17) Caillie, C. V.; Amos, R. D. *Chem. Phys. Lett.* **2000**, *317*, 159–164.
- (18) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433–7447.
- (19) Scalmani, G.; Frisch, M. J.; Mennucci, B.; Tomasi, J.; Cammi, R.; Barone, V. *J. Chem. Phys.* **2006**, *124*, 094107.
- (20) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947–12957.
- (21) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.
- (22) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.
- (23) Frauenheim, T.; Seifert, G.; Elstner, M.; Niehaus, T.; Kohler, C.; Amkreutz, M.; Sternberg, M.; Hajnal, Z.; Di Carlo, A.; Suhai, S. *J. Phys.: Condens. Matter* **2002**, *14*, 3015.
- (24) Koskinen, P.; Mäkinen, V. *Comput. Mater. Sci.* **2009**, *47*, 237–253.
- (25) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931–948.
- (26) Zheng, G.; Irle, S.; Morokuma, K. *Chem. Phys. Lett.* **2005**, *412*, 210–216.
- (27) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. A* **2006**, *110*, 13551–13559.
- (28) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751–5755.
- (29) Goyal, P.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2011**, *115*, 6790–6805.
- (30) Zheng, G.; Lundberg, M.; Jakowski, J.; Vreven, T.; Frisch, M. J.; Morokuma, K. *Int. J. Quantum Chem.* **2009**, *109*, 1841–1854.
- (31) Lundberg, M.; Sasakura, Y.; Zheng, G.; Morokuma, K. *J. Chem. Theory Comput.* **2010**, *6*, 1413–1427.
- (32) Trani, F.; Barone, V. *J. Chem. Theory Comput.* **2011**, *7*, 713–719.
- (33) Zheng, G.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, D. G.; Prabhakar, R.; Morokuma, K.; Lundberg, M.; Elstner, M.; Köhler, C.; Frauenheim, T. *J. Chem. Theory Comput.* **2007**, *3*, 1349–1367.
- (34) Moreira, N.; Dolgonos, G.; Aradi, B.; da Rosa, A.; Frauenheim, T. *J. Chem. Theory Comput.* **2009**, *5*, 605–614.
- (35) Dolgonos, G.; Aradi, B.; Moreira, N. H.; Frauenheim, T. *J. Chem. Theory Comput.* **2010**, *6*, 266–278.
- (36) Niehaus, T. A.; Suhai, S.; Della Sala, F.; Lugli, P.; Elstner, M.; Seifert, G.; Frauenheim, T. *Phys. Rev. B* **2001**, *63*, 085108.
- (37) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.
- (38) Kohler, C.; Seifert, G.; Gerstmann, U.; Elstner, M.; Overhof, H.; Frauenheim, T. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5109–5114.
- (39) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
- (40) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (41) Stewart, J. J. *Mol. Model.* **2007**, *13*, 1173–1213.
- (42) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.
- (43) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385.
- (44) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2010**, *6*, 1532–1537.
- (45) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C.; Valero, R.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2071–2085.
- (46) Silva-Junior, M. R.; Thiel, W. *J. Chem. Theory Comput.* **2010**, *6*, 1546–1564.
- (47) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.
- (48) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.

Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian Development Version*; Gaussian Inc.: Wallingford, CT, 2011.

(49) Barone, V.; Bloino, J.; Biczysko, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1092–1101.

On the Origin of the Chiro-Optical Activity in Supramolecular Assemblies: A Quantum Chemical Study of C_3 Octopolar Systems

Belén Nieto-Ortega,[†] Juan Casado,[†] Juan T. López Navarrete,[†] Gunter Henrich,[‡] and Francisco J. Ramírez^{*,†}

[†]Departamento de Química Física, Universidad de Málaga, 29071-Málaga, Spain

[‡]Departamento de Química Orgánica, Universidad Autónoma de Madrid, Cantoblanco, 28049-Madrid, Spain

S Supporting Information

ABSTRACT: First-principles quantum chemical approach has been used to understand the origin of the chiro-optical signal induced by the chiral aggregation of an achiral chromophore. The study was focused in predicting the circular dichroism (CD) spectra of different π -stacked columnar oligomers built with C_3 star-shape molecular bricks. We studied the influence of the relevant structural self-assembly parameters on the CD spectra (i.e., the number of units, the rotation angle and the intermonomer distance). A detailed analysis was based on the MO topologies and the magnetic and electric transition dipole moments, the vectors which determine the CD intensities, has been conducted. We have rationalized the influence of the various structural factors of supramolecular self-assemblies in connection with the nature of their CD spectroscopic signal, which provides new avenues for structure–spectroscopic relationships.

1. INTRODUCTION

Chiral molecular materials have become a major point of interest for scientists.^{1–4} Many of them are formed by organic molecules assembled into highly ordered helical superstructures in which they stack one above the other forming columns.^{5–7} This conformation gives interesting properties to the material, as the charge transportation along the column axis, useful in organic-based electronic devices.^{8,9} The presence of a stereogenic center in these molecules leads to the preferential formation of one of the two possible helices (enantiomeric excess) and to a noticeable enhancement of the ability to absorb left/right circularly polarized light differently;^{10–12} it is to say, the material is endowed with chiro-optical properties. The manipulation of these properties associated with the particular solid state structure provides a new avenue for applications.^{13–15} It turns out that the study of the supramolecular chirality is an essential step to develop new materials and future applications in diverse fields or organic electronics.

The chiro-optical spectroscopy (COS) has become a very useful tool to explore the structure of supramolecular assemblies, such as proteins, nucleic acids, or liquid crystals.^{16–20} As a first glance, the origin of the supramolecular chiro-optical signal is explained by the excitonic coupling theory between interacting chromophores.¹⁰ However, aspects as the growing of these three-dimensional (3D) structures and how COSs monitorize this process or the sensitivity of the chiro-optical signal with respect to the structural markets of the assemblies have not been thoroughly addressed up to now. Thus, while many molecular mechanics studies are routinely conducted to predict the helical behavior of chiral aggregates, first-principles quantum chemical studies addressing the same phenomena in aggregates are scarce to our knowledge. In this work we performed a detailed analysis of the electronic circular dichroism (CD) spectra, as predicted by quantum chemistry calculations, of several helical superstructures. As the molecular

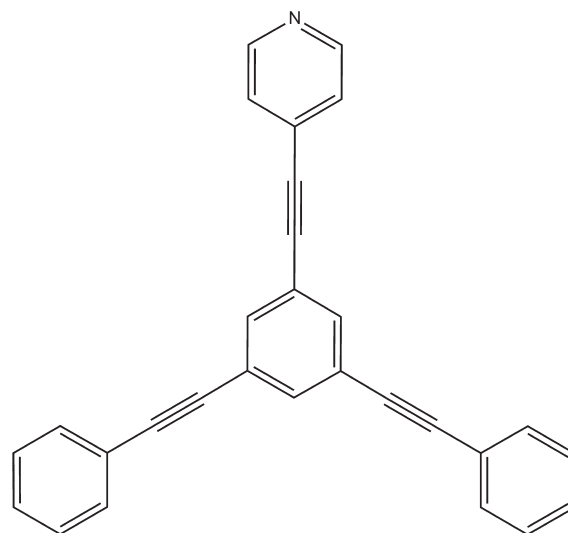


Figure 1. Molecular structure of the C_3 star-shaped octopolar molecular studied in this work.

brick we have chosen a C_3 octopolar molecule constituted by a central benzene ring substituted by three aryl-acetylene groups in alternating positions forming a star-shaped 2D configuration, Figure 1. A subtle structural difference in one of the three tails was included in order to control the rotation angle. Thus, two tails were phenyl-acetylene groups, while the third was a pyridine-acetylene one (Figure 1).

These C_3 aromatic planar structures can be easily functionalized by attaching aliphatic side chains containing asymmetric

Received: July 22, 2011

Published: September 02, 2011

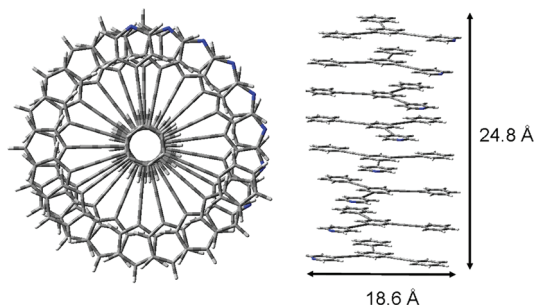


Figure 2. Front and side views of the columnar π -stacked octamer.

carbon atoms, which can act as stereogenic centers to promote chiral superstructures. From an experimental point of view, the stereocenters are effectively required to provide chiral activity to the material. They are not essential, nevertheless, in a theoretical study, as we can easily model any chiral superstructure from an achiral molecule as that displayed in the Figure 2. Nonetheless, the absence of intrinsic chirality in the molecules provides a significant advantage: The recorded chiral display will be fully induced by the assembly. It is then feasible and even advisable to provide results, at a theoretical level, about chiral assemblies of nonchiral molecules; this in turn will reduce the computational costs, thus allowing us to study more complex systems.

A variety of methods and strategies can be used to build a superstructure based on a C_3 star-shape molecule. Studies on these materials have been recently performed in our laboratory,²¹ providing experimental evidence that the main stabilizing force in C_3 -based supramolecular structure is the face-to-face π -stacking between the aromatic groups, which are spectroscopically well-defined achiral chromophores whose structures are often slightly perturbed when aliphatic side chains are attached. Following the previous report of Meijer et al.,^{22,23} we have focused our study on a right-handed helical columnar structure with a rotation angle of 20° . As shown in Figure 2 for one of the studied oligomers, these discotic structures allow for establishing strong π -stacking attractions and highly packed structures that can exhibit noticeable increases of their chiral features. The distance between monomers was established at 3.5 \AA , which is an intermediate value of the estimated optimal range for π -stacking interactions ($3.3\text{--}3.7 \text{ \AA}$).²⁴ The π -stacking strength was settled from a basic stabilization energy calculated from the following equation:

$$E_{\text{sta}} = \frac{E_{\text{olig}}}{n} - E_{\text{mon}}$$

(n = number of monomers in the oligomer)

The E_{sta} values, which have been included as Table TS1 of Supporting Information, increase with the oligomer size. We have then built a series of stable columnar oligomers, up to the octamer, able to generate nanometer fibers with a mean diameter of 18.6 \AA . The origin of the chiro-optical signal in these supramolecular systems was analyzed by predicting their CD spectra using DFT calculations in combination with ab initio methodology.

2. COMPUTATIONAL METHODS

The Gaussian'09 package of programs²⁵ was used for DFT quantum chemical calculations. The Becke's three parameter (B3)²⁶ gradient-corrected exchange functional was used, and the nonlocal correlation was provided by the Lee–Yang–Parr (LYP) expressions.^{27,28} To give account of charge-transfer

excitations, we used the Coulomb-attenuating hybrid method (CAM-B3LYP),²⁹ which includes the Hartree–Fock and the Becke exchanges as a variable ratio depending on the intermolecular distance. It has been demonstrated that this method gives an improved description of long-range interactions.^{30–35} Structural optimizations and spectroscopic features were calculated using the split-valence 3-21 g(d) basis set.^{36,37} At this level of calculation, data up to the octamer were available, while the essence of the chiro-optical signal was preserved. This was checked by comparing CD spectra for a right-handed trimer, with a rotation angle of 20° , at different ab initio methodologies, as 6-31 g(d,p) or cc-pvdz. These spectra are included as Figure FS1 of Supporting Information.

The minimum-energy structure of the monomer was achieved by allowing all the geometrical parameters to vary independently. Optimization gave us a fully planar star-shape structure in which the central benzene and the $C\equiv C$ bonds are insensible to the slight structural modification introduced in one of the three tails (the pyridine group), being all the benzene $C-C$ and the $C\equiv C$ distances of 1.398 and 1.205 \AA , respectively. Electronic excitation energies of the different oligomers were obtained by using the time-dependent DFT (TDDFT) formalism^{38,39} for which up to the 50 low-lying energy states were considered. The origin of the electronic chiral activity was explained at the light of the excitonic structure and the topology of the involved molecular orbitals (MO).

3. RESULTS AND DISCUSSION

The electronic absorption and CD spectra of the π -stacked columnar oligomers studied with rotation angle of 20° are shown in Figure 3. The intensity of all the spectra was divided by the number of monomers to remove the effect of the oligomer size. A first result is highlighted when comparing the panels of Figure 3: The two spectral series show opposite trends with respect to the oligomer size. Thus, while the intensity of the nonchiral signal decreases when increasing the number of oligomers, the chiral signal is clearly enhanced. The number of active bands also shows a different trend. The absorption spectra seem to show a single active band throughout the series whose wavelength monotonically decreases from the monomer to the octamer. However, no less than seven bands have appreciable CD intensities for the dimer, which are gradually removed when more C_3 units are stacked until showing only a bisignate, or Cotton band, near the wavelength of the absorption maximum of the octamer.

Figure 4 (left panel) shows the wavelength dispersion with respect to the number of monomers. The resulting trend suggests that an asymptote is almost reached for the octamer, from which we can deduce a blue-shift limit not greater than 10 nm in the case of an ideal infinite helix. The intensity changes with respect to the oligomer size, measured either on the absorption maxima or on the positive CD signals, are plotted in Figure 4 (right panel). This graphic contains a second and remarkable difference between the chiral and nonchiral spectra: Aside from the slope signs, only the absorption intensities show saturation when increasing the number of monomers, while the intensity of the CD signal shows a quasi-linear behavior, and no saturation is anticipated. Thus, these results highlight the sensitivity of the CD spectra to long-range interactions,^{40–42} as it gives account of the self-assembly even when the nonchiral signal is no longer able to.

If we defined the exciton bandwidth (W) as the energy difference between the bisignate maximum and minimum,⁴³ a relationship between W and the number of units of the helical

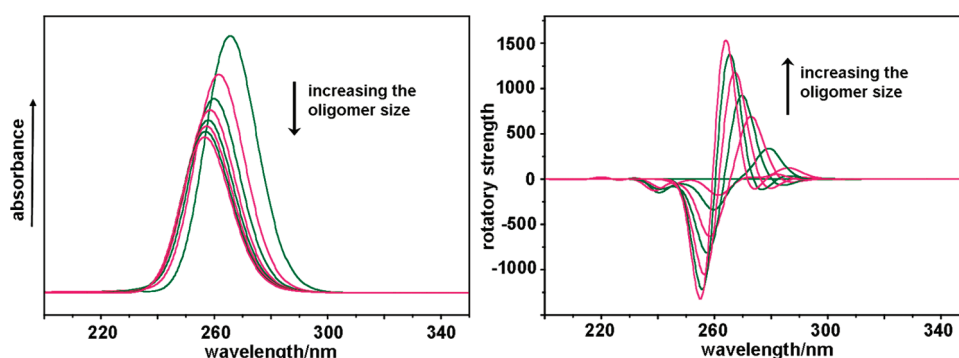


Figure 3. Electronic absorption (left panel) and CD (right panel) spectra of the π -stacked columnar oligomers, from one to eight monomers, with a rotation angle between stacked monomers of 20° . All the spectra were normalized to the number of monomers. Alternant colors are used to follow the spectral trend.

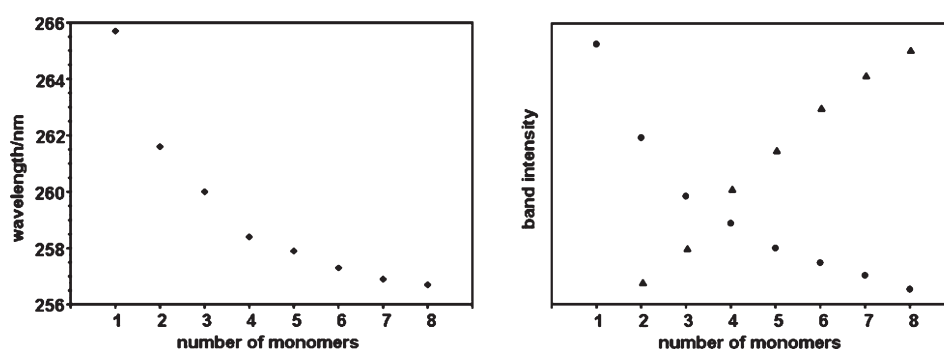


Figure 4. Left panel: graphic of the wavelength (at the absorption maxima) as a function of the number of monomers. Right panel: graphic of the intensity of the absorption band (circles) and the CD positive feature (triangles) as a function of the number of monomers.

Table 1. Exciton Bandwidth (W) Measured for the Main CD Bisignate of Each Oligomer

	W (meV)
dimer	430
trimer	340
tetramer	230
pentamer	220
hexamer	180
heptamer	150
octamer	110

assembly can be easily achieved. The calculated values for this series, Table 1, monotonically decrease when increasing the number of stacked units, thus demonstrating that W is useful in evaluating the interaction between the monomers.

The parameter W is also modulated by the distance between monomers (D). Accordingly with the following equation¹⁸ which relates D with the rotational strength (R):

$$R(\pm) = \pm \frac{\epsilon \mu^2 D}{4\hbar} \sin \alpha \sin \gamma \sin \tau$$

D is greater when R is increased. In this equation, τ is the rotation angle between monomers (here $+20^\circ$), while α and γ are the angles between the major helix axis and the electric transition dipole moments (ETDM) of two adjacent monomers. As will be discussed below, the ETDM vectors of the main transitions are

almost perpendicular to the helix axis, so that α and γ are near 1 (namely, $\alpha = 83.8^\circ$ and $\gamma = 82.5^\circ$ for the positive feature at 279 nm). This means that, in the oligomers studied, the rotational strength is largely dependent on the rotation angle τ .

The last result, nevertheless, would lead to the paradox that at $D = \infty$, the CD signal would have an infinite intensity. To compensate this effect, we have to consider the effect of D on the exciton bandwidth. This is given by the equation:¹⁸

$$W = \epsilon \pm \frac{\mu^2 (\sin \alpha \sin \gamma \cos \tau + 2 \cos \alpha \cos \gamma)}{D^3} \approx \epsilon \pm \frac{\mu^2 \cos \tau}{D^3}$$

where ϵ is the wavelength of the electronic transition in the monomer. As the fraction goes to 0 when $D \rightarrow \infty$, cancellation between the positive and the negative features of the bisignate will occur when increasing the intermonomer distance, and the Cotton effect disappears. We studied this phenomenon by calculating the CD spectra of the trimer at different D values, from 3 to 8 Å, which are shown in Figure 5. From the spectrum at 3 Å, a progressive increasing of the CD signal (R) together with a reduction of the exciton bandwidth (W) is predicted when increasing D up to 5 Å. However, and accordingly with the aforementioned equations, while W continues to decrease at 8 Å, the CD intensity is noticeably reduced as a consequence of the exciton cancellation. We have then obtained useful spectroscopic–structural relationships, as both the intensity of the CD signal and the width of the Cotton effect can be related, even at a quantitative level, with the intermonomer packing.

The results discussed above can be qualitatively interpreted from a molecular orbital analysis. We focus on the trimer as a simple stacked model. As can be seen in Figure 6 (left panels), the aforementioned bisignate is the envelope of four electronic transitions with excitation energies of 279.1 (D1) and 278.6 nm (D2) for the positive feature and 260.6 (D3) and 259.0 nm (D4) for the negative one. Among them, only D3 and D4 have nonvanished oscillator strength values, being the origin of the main absorption band in the UV–vis spectrum (Figure 6). These four excitations are described by the contribution of several mono-electronic transitions which involve up to 18 molecular orbitals, from MO289 to MO306, with MO297 being the highest occupied molecular orbital (HOMO). A complete summary of quantitative assignments for D1–D4 and topologies of the 289–306 MO are included as Table TS2 and Figure FS2, respectively, in Supporting Information. Figure 6 (right panels) also shows the calculated spectra for the trimer with a rotation angle of -20° , forming a left-handed helix. Its CD spectrum is a mirror image of the calculated one for the right-handed helix, while the absorption spectra are identical, which confirms that the chiro-optical signal is only due to the supramolecular assembly.

The similar optical activity of the two bisignate branches together with their extremely different absorption intensities allows us to anticipate that the positive CD feature should be the result of two magnetic dipole allowed (mda) transitions (D1 and D2) while the negative band would be due to two electric dipole

allowed (eda) ones. This conclusion is, therefore, depending of the angle θ between the corresponding magnetic and electric transition dipole moments (MTDM and ETDM, respectively), which determine the final CD activity of any electronic transition $\varphi_i \rightarrow \varphi_j$ by means of the Rosenfeld equation:¹⁸

$$\int \Delta \epsilon dv \propto R_{ij} \approx \vec{\mu}_{ij} \cdot \vec{m}_{ji} = |\vec{\mu}_{ij}| \cdot |\vec{m}_{ji}| \cdot \cos \theta$$

where R_{ij} is the rotational strength (which is related with the intensity of the chiro-optical signal), μ is the EDTM for the $\varphi_i \rightarrow \varphi_j$ transition, m is the MDTM for the reverse transition, and θ is the angle between these two magnetic dipole moments. As can be seen in the equation, the R_{ij} value will be positive if θ is smaller than 90° , negative if larger, and null if the vectors are perpendicular. In an achiral monomer, either $m = 0$, $\mu = 0$, or $\theta = 90^\circ$, so all its electronic transitions are CD silent, as successfully predicted in our case. However, a chiral assembly will modify these parameters, and as a consequence, an induced CD signal will appear.

Thus, to find out the true origin of the chiro-optical activity in the trimer (and by extension in the rest of oligomers), we need to calculate the EDTM and the MDTM vectors. They are drawn in Figure 7 for the D1–D4 electronic transitions. The components of each vector can be found in Table TS3 of Supporting Information. First we pay attention to D1 and D2 transitions. As represented in Figure 7, their calculated EDTM and MDTM vectors are within the xy molecular plane, being the θ values 2.4 and 4.2° , respectively. Hence the R_{ij} is positive. The modulus of their MDTMs is significantly greater than for their EDTMs, and as a consequence, we can conclude that D1 and D2 are markedly mda transitions with weak contributions from the EDTM.

The calculated MTDM and EDTM of D3 and D4 transition are also within the xy plane; however, their angles are rather different, 146.0 and 167.4° , respectively. The cosines of these angles are similar, in absolute values, to those of D1 and D2, but now the signs, and consequently the R_{ij} values, are negative. The MDTM modulus is here much lower than the EDTM one, what means that not only the sign of the band, but the origin of the CD signal is inverted. D3 and D4 are then eda transitions, that is to say, the nature of their chiro-optical activity is largely electric, having a weak contribution of the magnetic component. This is also the cause of their strong absorption intensities, which successfully explains the spectra of Figure 6.

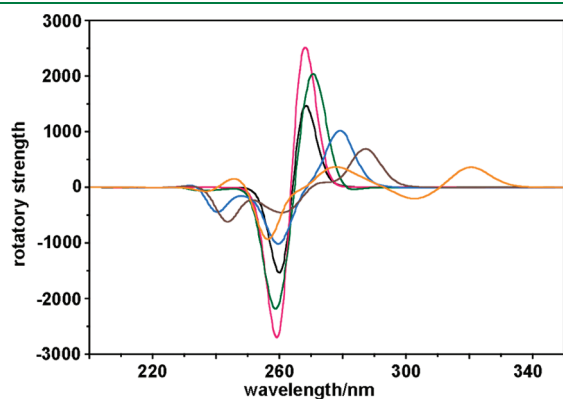


Figure 5. Calculated CD spectra of the right-handed (20° rotation angle) trimer at different intermolecular distances (D): 3.0 (orange), 3.3 (brown), 3.5 (blue), 4.0 (green), 5.0 (red), and 8.0 Å (black).

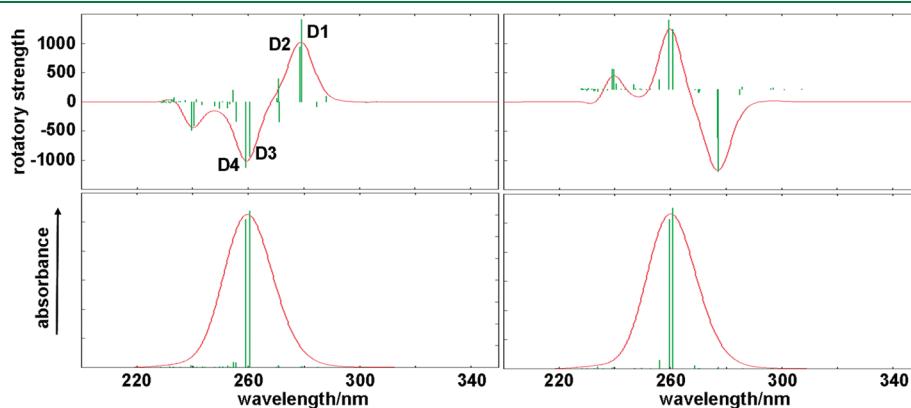


Figure 6. Bar representation of the exciton energies of the transitions calculated for the trimer with a rotation angle between stacked monomers of 20° (left panels) and -20° (right panels). The height of the green bars is proportional to the rotational strength.

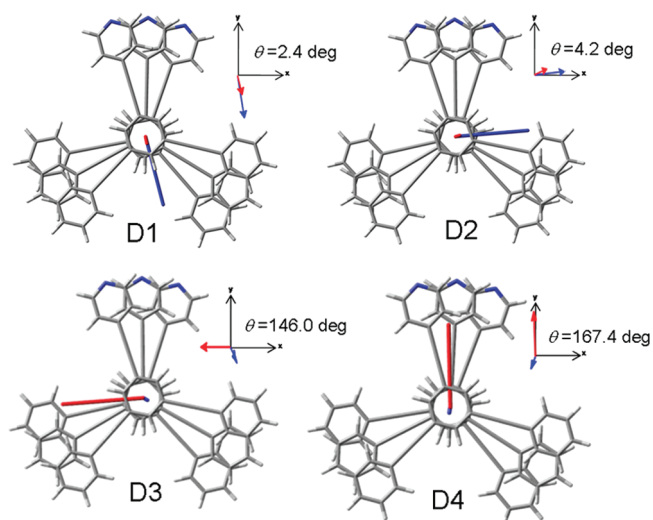


Figure 7. EDTM (red vectors) and MDTM (blue vectors) associated to D1–D4 electronic transitions of the trimer.

Table 2. Calculated Modulus of EDTM and MDTM Vectors Associated to the D1–D4 Monoelectron Transitions^a

	EDTM				MDTM			
	D1	D2	D3	D4	D1	D2	D3	D4
dimer	1.26	0.81	24.43	23.78	12.7	13.4	0.63	0.77
trimer	1.63	1.01	18.49	17.93	16.3	17.5	1.13	1.21
tetramer	2.09	1.17	12.65	12.93	19.3	21.3	1.64	1.46
pentamer	2.16	0.95	10.23	10.10	20.0	18.7	1.40	2.25
hexamer	2.27	0.97	9.77	9.20	21.4	21.8	1.52	2.22
heptamer	2.27	0.88	8.07	7.44	21.8	23.4	1.85	2.22
octamer	2.18	0.78	6.24	5.48	21.2	24.4	2.96	4.37

^a All values have been divided by the number of monomers in each case.

The MDTM and EDTM moduli of the electronic transitions D1–D4 allow us to explain the spectral intensities shown in Figure 3. Table 2 summarizes the values for the series of oligomers studied, all of them normalized to the number of monomers as performed for the calculated spectra. Accordingly with the behavior of the Cotton effect, the MDTM modulus grows up when increasing the oligomer size, showing similar values for the four transitions. On the contrary, the EDTM modulus is noticeably higher for D3 and D4 than for D1 and D2 doublets and, in agreement with the absorption spectra, the values for D3 and D4 go down with the number of monomers. Since EDTMs for D1 and D2 do not show a clear trend within the series, and taking into account that the θ values do not appreciably vary with respect to those shown in Figure 3 for the trimer, we can conclude that the increasing of the CD signal is largely caused by the enhancement of the magnetic coupling between the stacked monomers.

A further explanation of the different nature found for the two components of the Cotton effect in these oligomers can be reached by analyzing the MO of the trimer involved in the D1–D4 transitions. In an eda mechanism, only the displacement of electric charge during the excitation is relevant, while a mda transition is characterized by the magnitude of the magnetic moment. This means that in a mda transition, we must focus our

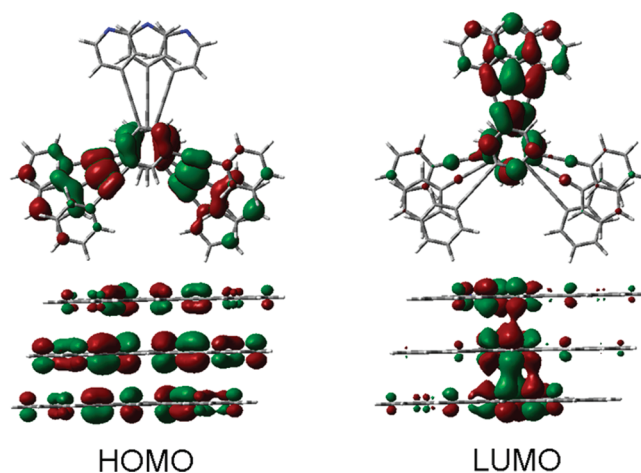


Figure 8. Frontal and side views of the HOMO and LUMO topologies calculated for the trimer.

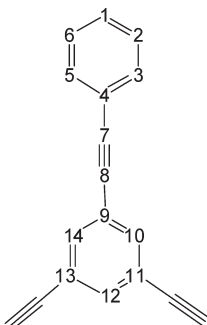
attention on the electric charge rotation perpendicular to the molecular planes, which can be evaluated in light of the MO topologies associated to the mono-electronic transitions and their relative weights (Figure FS2 and Table TS2, respectively, of Supporting Information). The analysis can be summarized in the following points:

- All MO analyzed belong to the π -bond system. In the basis of the electron density distribution, they can be classified into two groups. The 289, 291, 293, 298(LUMO), and 299 MOs have stacking topologies because the inter-plane electronic density is reinforced. The rest of MOs exhibit nodal planes in the intermonomer regions, so they are antistacking topologies. This fact can be observed for HOMO and LUMO, as representative examples, in Figure 8.
- The mono-electron transitions can connect two similar MO or not. Only in the last case we have charge rotation and, consequently, magnetic moment. This can be supported by calculating the total contribution of transitions with charge rotation in the two mda transitions, D1 and D2, being 23 and 35%, respectively, while these values are much lower in the two eda transitions, namely 3 and 7% for D3 and D4, respectively. These values also show the same trend as the calculated MDTM (see Table TS3, Supporting Information).
- As can be observed in Figure 8, and more extensively in Figure FS2, Supporting Information, in all the mono-electron transitions, the electronic density flows away from the $C\equiv C$ bonds. Electronic excitations are then associated to structural changes in which the $C\equiv C$ bonds become longer while the adjacent $C-C$ ones shorten, and as a consequence, the molecular structure in the excited state is expected to be more rigid than in the ground state. We expect this effect affects to the benzene rings, which undergo an aromatic-to-quinoid transition.

The last hypothesis has been investigated by comparing the optimized structures of the monomer in the ground state and in the first excited state, Table 3. They were calculated using the restricted configuration interaction (singlet) approach (RCIS) in combination with the 3-21 g(d) basis set. The quinoid structure is only observed for one of the three aryl-acetylene moiety and the central benzene ring, remaining the bond lengths of the other

Table 3. Relevant Bond Lengths (in Å) Calculated for the Ground State and the First Excited State of the Monomer

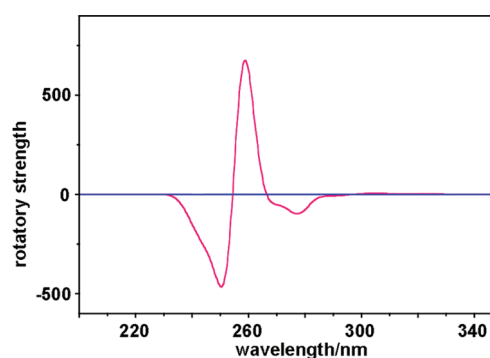
bond	S ₀	S ₁
C ₇ =C ₈	1.192	1.246
C ₄ -C ₇	1.432	1.371
C ₈ -C ₉	1.431	1.371
C ₂ =C ₃	1.382	1.369
C ₃ =C ₄	1.392	1.430
C ₁₀ =C ₁₁	1.389	1.376
C ₉ =C ₁₀	1.389	1.427



two tails unchanged. This is in agreement with the topology of most of the MO, in which the electronic density is concentrated over one of the aryl-acetylene groups. As predicted, the central ring distorts its hexagonal symmetry in the S₁ state, thus presenting a quinoid structure with respect to the involved aryl-acetylene group. Within this branch, the C≡C bond is lengthened, while the two adjacent C—C bonds are similarly shortened. This in turn modifies the outer benzene ring, which presented some quinoidization in the ground state, also in the aromatic-to-quinoid sense.

In the precedent paragraphs, we have shown that the CD signal produced by a 20° right-handed helices was largely due to either a magnetic or an electric mechanism. Looking for a relationship between the torsion angle of the chromophores and the chiro-optical response, we have studied a similar series of π -stacked columnar oligomers with a rotation angle between monomers of 120° and maintaining the intermonomer separation of 3.5 Å. Since this angle matches the pseudosymmetry angle of the C₃ molecules, the only structural element to characterize the helices is now the position of the pyridinium ring, without which no rotation angle could be defined, and in consequence, we would have no helix. In order to check the ability of the calculation method in this new scenario, we have first calculated the CD spectrum of a 120° right-handed tetramer in which the pyridine ring was replaced by a benzene one, that is to say, a columnar structure in which no rotation between monomers can be appreciated. Figure 9 compares the CD spectra of both tetramers, with or without pyridine. As can be seen, the chiral response in the last case is negligible, which demonstrates that our methodology is able to detect a helical symmetry even though it depends on a subtle atomic substitution in a moderate-size chromophore.

Figure 10 displays the normalized electronic absorption and CD spectra calculated for the new series. Comparing with the spectra of Figure 3, noticeable deviations are observed. While the UV–vis spectra of helices with 2–5 units are quite similar for both rotational angles, those with 6–8 units show a dramatic intensity reduction. The analysis of the CD spectra is more troublesome, as they do not exhibit a fair trend. A Cotton effect with a positive band at 259 nm and a negative one about 250 nm is the dominant feature in tetramer and pentamer, whose spectra also show very similar intensities after normalization. The pentamer shows a second, less intense, negative band at 271 nm, which undergoes intensity enhancement and blue-shift with the addition of monomers, being the only dominant band, at 266 nm, in the octamer. In a precedent paragraph, exciton cancellation has been related with the increasing of the intermonomer distance.

**Figure 9.** Calculated CD spectra of a π -stacked tetramer whose monomers contain either pyridine (red) or benzene (blue) moieties in one of their branches.

Now we show a similar phenomenon associated with the rotation angle between adjacent monomers, which is highlighted from a critical number of units in the helix.

The electronic transitions involved in the UV–vis and CD spectra of the tetramer (the smaller oligomer showing a clear Cotton effect) are analyzed in Figure 11, which also shows the associated MDTM and ECTM vectors of the most intense transitions. We name D1 and D2 to the positive and the negative strongest transitions, respectively. Figure 11 shows that the difference between the MDTM and EDTM moduli is less pronounced than in the case of the oligomers with rotation angle of 20° (see Figure 7), which indicates that no clear separation between magnetic and electric dipole mechanism exists. In consequence, chiro-optical activity is only predicted for the intense UV–vis absorption, while the positive CD features of the 20° oligomers, which were largely induced by the magnetic mechanism, do not have significant activity in this series. The topology and contribution of the MO associated to the D1 and D2 transitions (included as Figure FS3 and Table TS4 of the Supporting Information, respectively), indicate that in the more relevant mono-electron transitions, the π orbitals change their orientation from the xy plane to the z axis in both cases. This movement creates a charge rotation within the xy plane which explains both transitions are mda, besides being eda as supported by their strong absorption bands.

As the final part, we have applied all the arguments developed in this work to reach a suitable approach to the experimental CD spectrum of a C₃ star-shaped as a representative example of π -stacked columnar superstructures with chiral properties. As aforementioned, the presence of a stereogenic center is now mandatory to give chirality to the aggregates, so that we have selected the molecule shown in Figure 12, named as PA3Bz. It is formed by the basic aromatic skeleton studied in this work to which three identical aliphatic side chains, each of them containing an asymmetric carbon atom, were attached to the central benzene ring, with the outer pyridine moiety being replaced by a benzene ring.

In a previous work we reported the synthesis and a spectroscopic study of PA3Bz, supported by DFT theoretical calculations on a π -stacked dimer as a basic model of a left-handed helix.²¹ The experimental CD spectrum of a solid sample was largely composed by two broad features, namely a positive band at 271 nm and a negative one at 324 nm. In this spectral region we can take on that the CD signal is induced by the formation of chiral superstructures in solid state. By applying the results obtained in

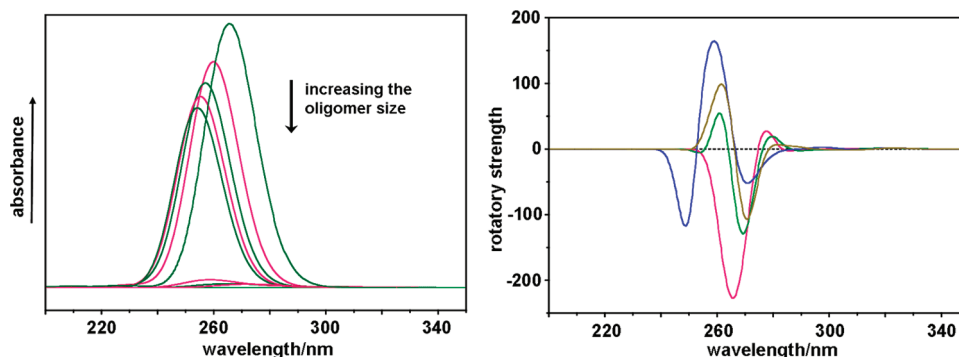


Figure 10. Left: Electronic absorption spectra of the π -stacked columnar oligomers, from 1 to 8 monomers, with a rotation angle between stacked monomers of 120° . Right: CD spectra of the oligomers with 5 (blue), 6 (brown), 7 (green), and 8 (red) monomers. All the spectra were normalized to the number of monomers.

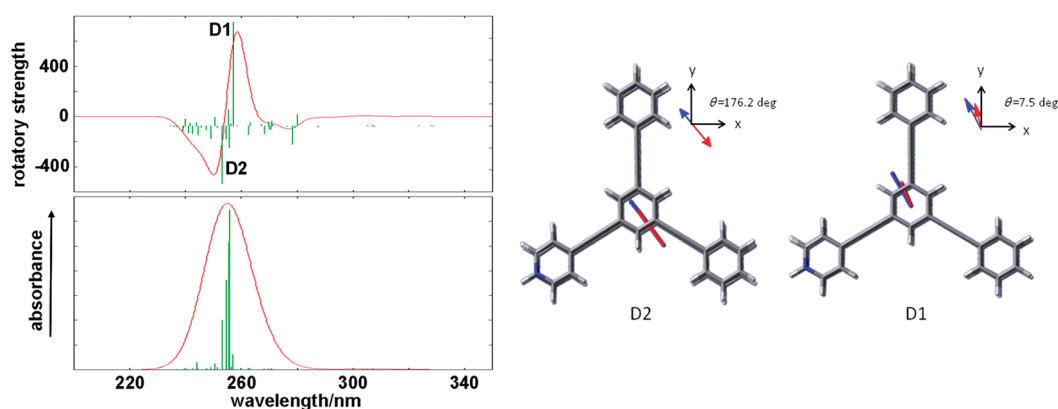


Figure 11. Left: bar representation of the exciton energies of the transitions calculated for a 120° right-handed tetramer. Right: EDTM (red) and MDTM (blue) vectors associated to the most intense electronic transitions.

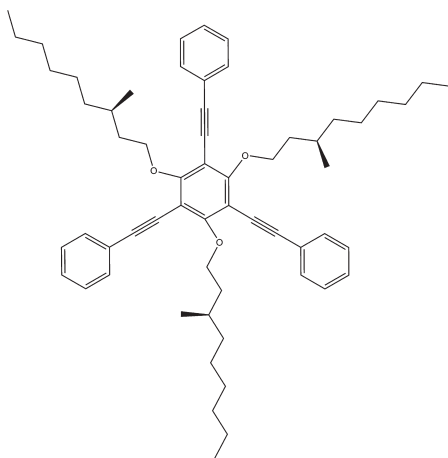


Figure 12. Chemical structure of the molecule PA3Bz.

this work for columnar helices, this experimental CD spectrum is compatible with a left-handed helix with a separation between stacked monomers of about $3.3\text{--}3.2\text{ \AA}$. In order to approach the rotation angle, we investigate the influence of this parameter on the chiro-optical signal. Figure 13 shows the CD spectra of a series of right-handed trimers with different rotation angles from

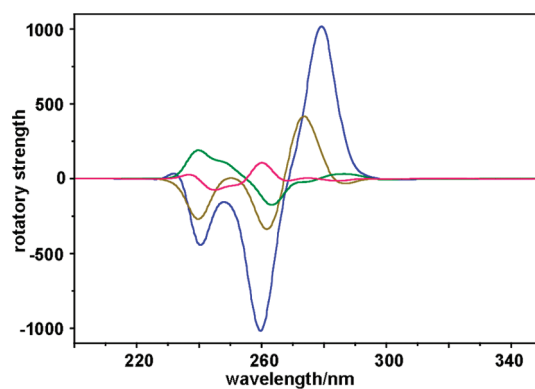


Figure 13. CD spectra of the right-handed trimer with rotation angles between monomers of 20° (blue), 40° (brown), 70° (green) and 120° (red).

20 to 120° and using the standard separation between monomers of 3.5 \AA . These spectra suggest that more intense Cotton effects are associated with small angles, being negligible the effect over the transition energies. We then chose a rotation angle of 20° to try to reproduce the experimental CD spectrum of PA3Bz.

Figure 14 shows a comparison between the experimental CD spectrum of PA3Bz from a solid sample and the calculated

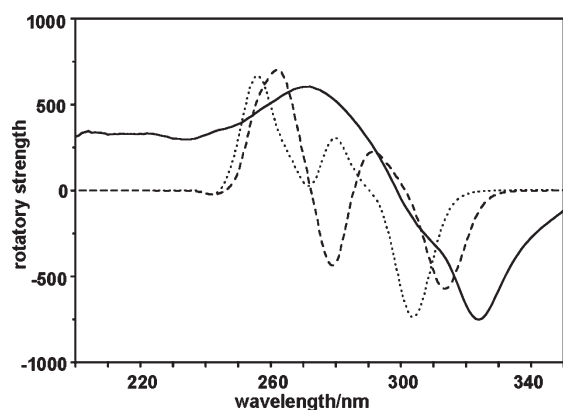


Figure 14. Comparison between the experimental CD spectrum of PA3Bz and calculated spectra using CAM-B3LYP/cc-pvdz at different intermonomer distances: 3.3 Å (dotted line) and 3.2 Å (broken line).

spectrum for the model trimer described in the previous paragraph, using the CAM-B3LYP functional in combination with the cc-pvdz ab initio basis set. The striking resemblance existing between these two spectra allows us to suggest that the model helix would be a reliable model of interaction in the solid state, at or at least partially, of the PA3Bz molecules.

4. CONCLUSIONS

We have applied a first-principles quantum chemical approach to understand the origin of the chiro-optical signal induced by the chiral aggregation of an achiral chromophore. The study was focused in predicting the CD spectra of different π -stacked columnar oligomers built with C_3 star-shape molecular bricks. We studied the influence of the relevant structural self-assembly parameters on the CD spectra (i.e., the number of units, the rotation angle, and the intermonomer distance). A detailed analysis was based on the MO topologies and the MTDM and ETDM, the vectors which determine the CD intensities, has been conducted.

We showed that the chiral signal per monomer unit (i.e., the total intensity divided by the number of monomers) increases almost linearly with increasing the oligomer size. The CD signal is also highly sensitive to long-range interunits distances. So, we have obtained useful spectroscopic–structural connections between this intermolecular distance and the exciton bandwidth. The analysis of the transition dipole moment vectors allowed us to accurately assign the magnetic and the electric contributions of the main CD bands, which are essential to explain the relations between the absorption and the CD spectra. The origin of the CD signal, either magnetic or electric, was successfully interpreted from of the topologies of the MOs associated to the one-electronic transitions. Two different MO topologies were described for this type of supramolecular aggregates, named as stacking and antistacking. Thus, from a simple analysis of the MOs connected in each CD band, its magnetic or electric character can be anticipated. Our calculations also showed that small rotation angles provide intense Cotton effects and strong CD bands. As a final remark, we rationale the influence of the various structural factors of supramolecular self-assemblies in connection with the nature of their CD spectroscopic signal, which provides new avenues for structure–spectroscopic relationships.

■ ASSOCIATED CONTENT

S Supporting Information. Details of stabilization energies of the oligomers, complete assignment of the more relevant excitation energies for the trimer, electric and magnetic transition dipole vectors, CD spectra using different ab initio basis sets, and topologies of the more relevant molecular orbitals. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ramirez@uma.es.

■ ACKNOWLEDGMENT

Financial support is indebted to DGES-MEC and Junta de Andalucía through research projects CTQ2009-10098-BQU and P09-FQM-4708. B.N.O. thanks the Ministerio de Educación (Spain) for the personal grant FPU-AP2009-2797.

■ REFERENCES

- (1) Seeber, G.; Tiedemann, B. E. F.; Raymond, K. N. In *Topics in Current Chemistry*; Crego-Calama, M., Reinhoudt, D. N., Eds.; Springer: Berlin, Germany, 2006; Vol. 265, Chpt. 4.
- (2) Soai, K.; Kawasaki, T. In *Topics in Current Chemistry*; Soai, K., Ed.; Springer: Berlin, Germany, 2008; Vol. 284, Chpt. 1.
- (3) Raval, R. *Chem. Soc. Rev.* **2009**, *38*, 707–721.
- (4) Crassous, J. *Chem. Soc. Rev.* **2009**, *38*, 830–845.
- (5) Hoeben, F. J. M.; Jonkheijm, P.; Meijer, E. W.; Schenning, A. P. H. *J. Chem. Rev.* **2005**, *105*, 1491–1546.
- (6) Pisula, W.; Tomovic, Z.; Watson, M. D.; Müllen, K.; Kussmann, J.; Ochsenfeld, C.; Metzroth, T.; Gauss, J. *J. Phys. Chem. B* **2007**, *111*, 7481–7487.
- (7) (a) J.-M. Lehn. *Supramolecular Chemistry: Concepts and Perspectives*, Wiley-VCH: Weinheim, Germany, 1995. (b) Lehn, J.-M. *Science* **2002**, *295*, 2400–2403.
- (8) Meijer, E. W.; Schenning, Albert P. H. *J. Nature* **2002**, *419*, 353–354.
- (9) Coquerel, G.; Amabilino, D. B. In *Chirality at the Nanoscale: Nanoparticles, Surfaces, Materials and more*; D. B. Amabilino, D. B., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2009.
- (10) Sznatke, G. In *Circular Dichroism. Principles and Applications*; Berova, N.; Nakanish, K.; Woody, R. W., Eds; Wiley-VCH: New York, 2000; Chpt. 1.
- (11) Gawronski, J., Skowronek, P. In *Chiral Analysis*; Busch, K. W.; Busch, M. A., Eds. Elsevier B.V: Amsterdam, The Netherlands, 2006; Chpt. 13.
- (12) Nordén, B.; Rodger, A.; Dafford, T. *Linear Dichroism and Circular Dichroism*; RSC Publishing: Cambridge, U.K., 2010; Chpt. 10.
- (13) Kuroda, R.; Honma, T. *Chirality* **2000**, *12*, 269–277.
- (14) Johannessena, C.; Thulstrup, P. W. *Dalton Trans.* **2007**, 1028–1033.
- (15) Castiglioni, E.; Biscarini, P.; Abbate, S. *Chirality* **2009**, *21*, E28–E36.
- (16) *Circular Dichroism and the Conformational Analysis of Biomolecules*; Fasman, G. D., Ed.; Plenum Press: New York, 1996.
- (17) Lightner, D. A.; Gurst, J. E. *Organic Conformational Analysis and Stereochemistry from Circular Dichroism Spectroscopy*; Wiley-VCH: New York, 2000.
- (18) Nordén, B.; Rodger, A.; Dafforn, T. *Linear Dichroism and Circular Dichroism*; RSC Publishing: Cambridge, U.K., 2010; Chpt. 8.
- (19) Gottarelli, G.; Lena, S.; Masiero, S.; Pieraccini, S.; Spada, G. P. *Chirality* **2008**, *20*, 471–485.
- (20) Lee, C. C.; Grenier, C.; Meijer, E. W.; Schenning, A. P. H. *J. Chem. Soc. Rev.* **2009**, *38*, 671–683.
- (21) Nieto-Ortega, B.; Ramirez, F. J.; Hennrich, G.; Gomez-Lor, B.; Casado, J.; Lopez Navarrete, J. T. *J. Phys. Chem. B* **2010**, *114*, 5710–5717.

- (22) Smulders, M. M. J.; Stals, P. J. M.; Mes, T.; Paffen, T. F. E.; Schenning, A. P. H. J.; Palmans, A. R. A.; Meijer, E. W. *J. Am. Chem. Soc.* **2010**, *132*, 621–626.
- (23) van Gorp, J. J.; Vekemans, J. A. J. M.; Meijer, E. W. *J. Am. Chem. Soc.* **2002**, *124*, 14759–14769.
- (24) Steed, J. W.; Atwood, J. L. *Supramolecular Chemistry*, Wiley: New York, 2009, Chpt. 10.
- (25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.2; Gaussian, Inc.: Wallingford, CT, 2009.
- (26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (27) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (28) Kim, K.; Jordan, K. D. *J. Phys. Chem.* **1994**, *98*, 10089–10094.
- (29) Yanai, T.; Tew, D. P.; Nicholas, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (30) Peach, M. J. G.; Tellgren, E. I.; Salek, P.; Helgaker, T.; Tozer, D. J. *J. Phys. Chem. A* **2007**, *111*, 11930–11935.
- (31) Limacher, P. A.; Mikkelsen, K. V.; L_{thi}, H. P. *J. Chem. Phys.* **2009**, *130*, 194114/1–194114/7.
- (32) Jacquemin, D.; Wathélet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.
- (33) Zhang, J.; Guo, X.; Cao, Z. *J. Chem. Phys.* **2009**, *131*, 144307/1–144307/7.
- (34) Borini, S.; Limacher, P. A.; Lüthi, H. P. *J. Phys. Chem. A* **2010**, *114*, 2221–2229.
- (35) Rivera- Rivera-Fuentes, P.; Alonso-Gómez, J. L.; Petrovic, A. G.; Seiler, P.; Santoro, F.; Harada, N.; Berova, N.; Rzepa, H. S.; Diederich, F. *Chem. Eur. J. A* **2010**, *16*, 9796–9807.
- (36) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (37) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. *J. Comput. Chem.* **1983**, *4*, 294–301.
- (38) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997–1000.
- (39) Gross, E. K. U.; Kohn, W. *Adv. Quantum Chem.* **1990**, *21*, 255–294.
- (40) Matile, S.; Berova, N.; Nakanishi, K.; Fleischhauer, J.; Woody, R. W. *J. Am. Chem. Soc.* **1996**, *118*, 5198–5206.
- (41) Lewis, F. D.; Liu, X. Y.; Wu, Y. S.; Zuo, X. B. *J. Am. Chem. Soc.* **2003**, *125*, 12729–12731.
- (42) Tsubaki, K.; Takaishi, K.; Tanaka, H.; Miura, M.; Kawabata, T. *Org. Lett.* **2006**, *8*, 2587–2590.
- (43) Van Dijk, L.; Bobbert, P. A.; Spano, F. C. *J. Phys. Chem. B* **2010**, *114*, 817–825.

Basis Set Dependence of Vibrational Raman and Raman Optical Activity Intensities

James R. Cheeseman* and Michael J. Frisch

Gaussian Inc., 340 Quinipiac St., Bldg. 40, Wallingford, Connecticut 06492-4050, United States

ABSTRACT: We present a systematic study of the basis set dependence of the backscattering vibrational Raman intensities and Raman Optical Activity (ROA) intensity differences. The accuracies of computed Raman intensities and ROA intensity differences for a series of commonly used basis sets are reported, relative to large reference basis sets, using the B3LYP density functional. This study attempts to separately quantify the relative accuracies obtained from particular basis set combinations: one for the geometry optimization and force field computation and the other for the computation of Raman and ROA tensors. We demonstrate here that the basis set requirements for the geometry and force fields are not similar to those of the Raman and ROA tensors. The Raman and ROA tensors require basis sets with diffuse functions, while geometry optimizations and force field computations typically do not. Eleven molecules were examined: (S)-methyloxirane, (S)-methylthirane, (R)-epichlorhydrin, (S)-CHFClBr, (1S,5S)- α -pinene, (1S,5S)- β -pinene, (1S,4S)-norborneneone, (M)- σ -[4]-helicene, an enone precursor to a cytotoxic sesquiterpene, the gauche-gauche conformer of the monosaccharide methyl- β -D-glucopyranose, and the dipeptide Ac-(alanine)₂-NH₂. For the molecules examined here, intensities and intensity differences obtained from Raman and ROA tensors computed using the aug-cc-pVDZ basis set are nearly equivalent to those computed with the larger aug-cc-pVTZ basis set. We find that modifying the aug-cc-pVDZ basis set by removing the set of diffuse d functions on all atoms (while keeping the diffuse s and p sets), denoted as aug(sp)-cc-pVDZ, results in a basis set which is significantly faster without much reduction in the overall accuracy. In addition, the popular rDPS basis set introduced by Zuber and Hug offers a good compromise between accuracy and efficiency. The combination of either the aug(sp)-pVDZ or rDPS basis for the computation of the Raman and ROA tensors with the 6-31G* basis set for the geometry optimization and force field calculation is a reliable and cost-effective method for obtaining Raman intensities and ROA intensity differences.

1. INTRODUCTION

Raman optical activity (ROA), the difference in Raman scattering intensity for left and right circularly polarized light, is a powerful tool for studying chiral molecules and determining the structure of biomolecules in their native environment. The theory and first genuine experimental observations of ROA were reported by Barron and Buckingham,^{1,2} in the early 1970s. Recent advances in instrumentation³ and the development of a commercial ROA instrument (the ChiralRAMAN from Bio-Tools, Inc.) have led to an increase in the number of applications. See the review by Barron and Buckingham⁴ for the current status of the field. The calculation of ROA spectra using *ab initio* quantum chemical methods is an important aspect of the technique, as comparison of predicted and experimental spectra can provide detailed structural information as well as absolute configuration assignments. There have been many recent computational applications of ROA for a variety of systems, including amino acids,^{5,6} small peptides,^{6–11} proteins,¹² transition metal complexes,¹³ carbohydrates,^{14–16} and helicenes^{17,18} as well as many other molecules.^{19–23} ROA has been used to study atropisomerism in binaphthyl derivatives,²⁴ and recently, the absolute configuration of junonone, a natural monoterpene, was determined using ROA.²⁵ See ref 26 for a review of recent computational applications.

There are several forms of ROA depending on the choice of polarization modulation, scattering geometry, and laser frequency.^{27–32} In the far-from-resonance theory, where the exciting laser radiation is far from the lowest allowed electronic excited state, ROA intensity differences depend on the normal mode derivatives of three polarizability tensors, namely, the

electric dipole–electric dipole polarizability, the electric dipole–electric quadrupole polarizability, and the electric dipole–magnetic dipole polarizability. The electric dipole–electric dipole polarizability derivatives are used to form two Raman tensor invariants, and all three polarizability derivatives are used to form the three ROA tensor invariants^{31,33} from which the Raman intensities and ROA intensity differences are obtained, respectively.

The recent development of fully analytic derivative methods to compute the various polarizability derivatives^{34–37} has further extended the size of systems that can be studied. See ref 33 for a review of theoretical approaches for computing ROA intensities. Fully analytic methods offer two choices:³³ a “one-step” procedure in which the force field and Raman/ROA tensor invariants are computed at the same time using the same level of theory and taking advantage of the 2n+1 rule to do only first order CPKS or a “two-step” procedure in which the force field and Raman/ROA tensor invariants are computed in separate steps and require second order CPKS (n+1 rule), thus allowing for the use of different levels of theory for each step. Both procedures require the solution of 11 frequency-dependent first order CPKS equations (three electric, three magnetic, and five quadrupole fields) as well as three times the number of atoms static first order nuclear coordinate CPKS equations. The “two-step” procedure additionally requires the solution of 30 static mixed second order CPKS equations (six electric–electric, nine electric–magnetic, 15 electric–quadrupole). (We note that in Furche’s Lagrangian

Received: July 22, 2011

Published: September 13, 2011

approach for computing Raman intensities,³⁸ the static second order CPKS equations are referred to as “TDKS gradient equations”, but the computations involved are identical.) If the basis set requirements for the force field and Raman/ROA tensor invariants are similar, then the “one-step” procedure is optimal, since the additional solution of 30 static CPKS equations is avoided. We note that it is possible to use different levels of theory for the force field and Raman/ROA tensor invariants in an effective “one-step” procedure where the resulting nuclear coordinate derivatives of the three polarizability tensors are combined with a force field computed at a different level of theory. In this case, however, three times the number of atoms static nuclear coordinate CPKS equations need to be solved for each level of theory. Since the number of nuclear coordinate CPKS equations that need to be solved depends on the number of atoms, this step becomes dominant for large molecules. All previous published computational ROA studies have been carried out using a “one-step” procedure or an “effective one-step” procedure as outlined above. Although not addressed in this work, the “two-step” ($n+1$ rule) has the additional advantage that it allows for alternative approaches to full force-field determination such as mode-tracking^{23,39} and intensity-tracking^{40,41} methodologies developed by Reiher and co-workers.

Raman intensities are known to be sensitive to diffuse augmentation of the basis set.^{38,42} Zuber and Hug⁴³ have pointed out the importance of diffuse functions in computing ROA tensor invariants. They proposed a minimal rDPS basis set, which is the 3-21++G basis augmented with a semidiffuse p function with an exponent of 0.2 on hydrogen, which provided ROA intensity differences which were close to those obtained using the aug-cc-pVDZ basis set. Although ROA tensor invariants computed using many basis sets were compared in this study, only one basis set was used for the force field. Another basis set study of ROA intensity differences has previously been presented by Ruud and Reiher.⁴⁴ They observed comparably good results for the aug-cc-pVDZ and aug-cc-pVTZ basis sets and concluded that the small rDPS basis set proposed by Zuber and Hug⁴³ was able to reproduce the ROA intensity differences with sufficient accuracy. With the exception of rDPS, ROA intensity differences were obtained from ROA tensors and force fields which were computed using the same basis set. Neither of these studies examined the dependence of the ROA intensity differences on the level of theory used to compute the optimized geometry and force field. These previous studies also used a numerical procedure for evaluating the nuclear coordinate derivatives of the three polarizability tensors—and were therefore limited to small molecules due to the very demanding computational requirements.

In addition to the quality of the basis set used to compute the Raman/ROA tensor invariants, Raman intensities and ROA intensity differences are also dependent on the quality of the force field via the transformation to normal coordinates. The goal of this study is to separate and quantify these two effects through a systematic study of basis set dependence of the Raman intensities and ROA intensity differences. We examine the dependence of the Raman intensities and ROA intensity differences on the basis set used to compute the Raman/ROA tensors invariants as well as the basis set used for the geometry optimization and force field determination. We also report the relative accuracies of various basis set combinations and address which combinations provide the best accuracy for the cost. Since the purpose of this study is to examine and quantify the basis set

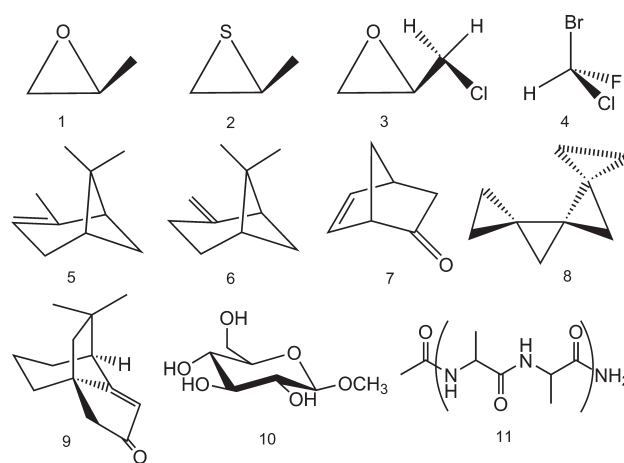


Figure 1. Molecular structures for (S)-methyloxirane (**1**), (S)-methylthirane (**2**), (R)-epichlorhydrin (**3**), (S)-CHFClBr (**4**), (1S,5S)- α -pinene (**5**), (1S,5S)- β -pinene (**6**), (M)- σ -[4]-helicene (**7**), (1S,4S)-norborneneone (**8**), an enone precursor to the cytotoxic sesquiterpene nor-suberosenone (1R,11S)-**9**, methyl- β -D-glucopyranose (**10**), and Ac-(alanine)₂-NH₂ (**11**).

convergence of Raman intensities and ROA intensity differences, one functional (B3LYP) is used throughout. Solvent effects and anharmonic corrections are also ignored.

2. METHODS

The molecules chosen for this study, shown in Figure 1, are (S)-methyloxirane (**1**), (S)-methylthirane (**2**), (R)-epichlorhydrin (lowest energy conformer) (**3**), (S)-CHFClBr (**4**), (1S,5S)- α -pinene (**5**), (1S,5S)- β -pinene (**6**), (M)- σ -[4]-helicene (**7**), (1S,4S)-norborneneone (**8**), an enone precursor to the cytotoxic sesquiterpene nor-suberosenone (1R,11S)-**9**, methyl- β -D-glucopyranose (gauche–gauche conformer) (**10**), and Ac-(alanine)₂-NH₂ (**11**). Molecules **1–3** were chosen because of their small size, thus allowing the use of the large aug-cc-pVQZ basis set as a reference. In addition, they include the second row atoms S and Cl. The ROA spectra has previously been studied, both computationally and experimentally for **1**^{44,45} and computationally for **3**.^{44,46} Molecule **4** is one of the simplest examples of the (Le Bel and van't Hoff) asymmetric carbon atom. Polavarapu et al.⁴⁷ provided the assignment of the absolute configuration on the basis of an *ab initio* ROA spectra calculation. It is included in this study in order to examine the requirement of higher angular momentum diffuse functions. Backscattering ROA measurements⁴⁸ and calculated ROA spectra⁴⁹ have been previously presented for **5**. For **6**, both forward scattering ROA⁵⁰ and backscattering ROA⁵¹ measurements as well as calculated ROA spectra⁴⁶ have been presented. Molecule **7** is potentially of interest given the very large value of specific rotation at the sodium D line.^{52,53} The ROA spectra of **8** have been measured by Hug et al.,¹⁷ and previous calculations^{44,46} have also been presented. Molecule **9**, a chiral enone precursor to the cytotoxic sesquiterpene nor-suberosenone and nor-suberosanone, was synthesized by Jean-Charles and co-workers⁵⁴ and its absolute configuration determined by Stephens and co-workers⁵⁵ using VCD. There have been several recent ROA calculations presented for sugars^{15,16,56} and carbohydrates.¹⁴ The experimental and calculated Raman and ROA spectra of **10** have been presented previously.^{56,57} Given the many recent ROA calculations presented for peptides,^{6–12} molecule **11** was chosen as an example peptide model. We note

that eight of the molecules examined here (1, 2, 4–9) are conformationally rigid, and therefore only one conformation is present at room temperature.

The expression for the SCP backscattered Raman intensities ($I^R + I^L$) and ROA intensity differences ($I^R - I^L$) are given by eq 1:

$$I^R + I^L \propto \frac{(\tilde{\nu}_{\text{in}} - \tilde{\nu}_i)^4}{1 - \exp[-hc\tilde{\nu}_i/(k_B T)]} [45\alpha_i^2 + 7\beta_i^2]$$

$$I^R - I^L \propto \frac{(\tilde{\nu}_{\text{in}} - \tilde{\nu}_i)^4}{1 - \exp[-hc\tilde{\nu}_i/(k_B T)]} [12\beta_{G_i}^2 + 4\beta_{A_i}^2] \quad (1)$$

where

$$\alpha_i^2 = \frac{1}{9} \left(\frac{\partial \alpha_{\alpha\alpha}}{\partial Q_i} \right)_0 \left(\frac{\partial \alpha_{\beta\beta}}{\partial Q_i} \right)_0$$

$$\beta_i^2 = \frac{1}{2} \left[3 \left(\frac{\partial \alpha_{\alpha\beta}}{\partial Q_i} \right)_0 \left(\frac{\partial \alpha_{\alpha\beta}}{\partial Q_i} \right)_0 - \left(\frac{\partial \alpha_{\alpha\alpha}}{\partial Q_i} \right)_0 \left(\frac{\partial \alpha_{\beta\beta}}{\partial Q_i} \right)_0 \right]$$

$$\beta_{G_i}^2 = \frac{1}{2} \left[3 \left(\frac{\partial \alpha_{\alpha\beta}}{\partial Q_i} \right)_0 \left(\frac{\partial G'_{\alpha\beta}}{\partial Q_i} \right)_0 - \left(\frac{\partial \alpha_{\alpha\alpha}}{\partial Q_i} \right)_0 \left(\frac{\partial G'_{\beta\beta}}{\partial Q_i} \right)_0 \right]$$

$$\beta_{A_i}^2 = \frac{\omega}{2} \left[\left(\frac{\partial \alpha_{\alpha\beta}}{\partial Q_p} \right)_0 \left(\frac{\partial \varepsilon_{\alpha\gamma} \delta \delta A_{\gamma, \delta\beta}}{\partial Q_p} \right)_0 \right] \quad (2)$$

In eqs 1 and 2, α_i^2 and β_i^2 are the Raman invariants, $\beta_{G_i}^2$ and $\beta_{A_i}^2$ are the ROA invariants, and $\tilde{\nu}_{\text{in}}$ and $\tilde{\nu}_i$ are the wave numbers of the incident light (532 nm) and of the i th vibrational mode, respectively. Q_i is the normal mode of the i th vibration, and $\alpha_{\alpha\beta}$, $G'_{\alpha\beta}$, and $A_{\alpha\beta\gamma}$ are the frequency dependent electric dipole–electric dipole, electric dipole–magnetic dipole, and electric dipole–electric quadrupole polarizabilities, respectively. Equation 1 includes the ν^4 and Boltzmann factors—which are necessary for comparing calculated spectra to experimental spectra.^{22,28,31,44,58,59} Since absolute intensities and intensity differences are not typically measured experimentally, calculated Raman intensities and ROA intensity differences are therefore compared in arbitrary units. We use percent normalized RMS deviations (%NRMS), which is the RMS deviation divided by the range of the reference. For the four smallest molecules (1–4), the aug-cc-pVQZ basis set was used as a reference for the geometry, force field, Raman, and ROA tensors. For the other molecules (5–11), the aug-cc-pVTZ basis set was used as a reference for the geometry, force field, and Raman and ROA tensors. We examine the number of wrong signs, within 10% of the range. This provides a more meaningful measure of the quality as only significant wrong signs are seen, thus eliminating small, insignificant values.

3. COMPUTATIONAL DETAILS

All calculations were performed using a development version of Gaussian.⁶⁰ The B3LYP functional was used throughout. Frequency dependent ROA and Raman tensors were computed using magnetic field dependent basis functions (GIAOs)^{61,62} using either the “one-step” (2n+1) or “two-step” procedure (n+1) algorithm as described by Ruud and Thorvaldsen.³³ The incident light frequency was 532 nm. For the “one-step” procedure, geometry optimization, force field, and Raman and ROA tensors were computed using the same level of theory. For the

“two-step” procedure, geometry optimization and the force field computation were carried out in the first step while Raman and ROA tensors were computed in the second step using the geometry and force field obtained from the first step via the checkpoint file. Geometry optimizations used tight convergence criteria where the maximum force was less than 1×10^{-5} au. A pruned (99, 590) grid having 99 radial shells and 590 angular points per shell was used throughout. Raman intensities and ROA intensity differences were computed from the appropriate SCP backscattering combinations of tensor invariants^{28,31,33} as in eq 1. Intensities, in arbitrary units, are plotted using a Lorentzian line shape with a half width of 10 cm^{-1} . Percent normalized RMS (%NRMS) is the RMS deviation divided by the range of the reference, multiplied by 100. Modes below 100 cm^{-1} were excluded from the analysis. NWS are the number of wrong signs which have intensity differences ($I^R - I^L$) greater than 10% of the total range of the reference. Tables in the paper include vibrational frequencies in the range $100\text{--}1900 \text{ cm}^{-1}$. Differences in the force fields may lead to a different ordering of the vibrational modes, giving rise to apparent differences in the sign of the intensity differences as well as to artificially large %NRMS values. This was particularly the case for the 6-31G* and aug-cc-pVDZ force field basis sets. Care was taken to match corresponding modes, in terms of atomic displacements, in the comparison.

The basis sets used for geometry optimization and force field calculation were aug-cc-pVQZ (molecules 1–4 only), aug-cc-pVTZ, cc-pVTZ, 6-31G*, aug-cc-pVDZ, and aug(sp)-cc-pVDZ. The basis sets used to compute Raman/ROA tensors were aug-cc-pVQZ (molecules 1–4 only), aug-cc-pVTZ, aug-cc-pVDZ, aug(sp)-cc-pVDZ, rDPS, cc-pVQZ (molecules 1–4 only), cc-pVTZ, cc-pVDZ, 6-311++G(2d,2p), 6-311++G**, 6-31G*, TZVP, d-aug-cc-pVTZ, d-aug-cc-pVDZ, and Sadlej and LPol-ds. We use the common convention where the basis set listed before “//” implies the one which was used to compute the Raman and ROA tensors, while the basis set listed after “//” implies the one which was used for the geometry minimization and force field determination. For molecules 1–4, the aug-cc-pVQZ//aug-cc-pVQZ level of theory was used as the reference, and for molecules 5–11, the aug-cc-pVTZ//aug-cc-pVTZ level of theory was used as the reference. For each of the molecules, the numbers of basis functions for many of the basis sets are given in Table 1, along with the number of normal modes considered in the analysis.

The basis set references are as follows. Pople style basis sets: 6-311++G(2d,2p), 6-311++G**, and 6-31G*.^{63–70} Dunning’s correlation consistent basis sets: cc-pVDZ, cc-pVTZ, and cc-pVQZ.⁷¹ Dunning’s singly augmented correlation consistent basis sets: aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ.⁷² The d-aug-cc-pVTZ and d-aug-cc-pVDZ⁷³ are the doubly augmented aug-cc-pVTZ and aug-cc-pVDZ basis sets, respectively. Sadlej basis sets: Sadlej^{74,75} and LPol-ds.⁷⁶ Ahlrichs’ TZVP^{77,78} basis set. The rDPS basis set⁴³ is the 3-21++G basis set augmented with a semidiffuse p function with an exponent of 0.2 on hydrogen. The aug(sp)-cc-pVDZ basis set, introduced here, is obtained from the aug-cc-pVDZ basis set by removing all d type diffuse functions, leaving only s and p type diffuse functions. We note the previous use of pruned augmented basis sets for computing magnetizabilities.⁷⁹

4. RESULTS AND DISCUSSION

4.1. Molecules 1–3. The %NRMS deviations for the frequencies, Raman intensities and ROA intensity differences, and

Table 1. Number of Basis Functions and Normal Modes for Molecules 1–11

molecule	stoichiometry	modes ^a	aug-cc-pVQZ	cc-pVQZ	aug-cc-pVTZ	cc-pVTZ	aug-cc-pVDZ	aug(sp)-cc-pVDZ	rDPS	6-31G*
(1) (methyloxirane)	C ₃ H ₆ O	18	596	400	322	204	146	126	88	72
(2) (methylthirane)	C ₃ H ₆ S	18	600	404	326	208	150	121	92	76
(3) (epichlorhydrin)	C ₃ H ₅ ClO	18	634	429	349	224	164	139	99	89
(4) (CHFClBr)	CHFClBr	8	383	267	224	151	118	98	75	81
(5) (α -pinene)	C ₁₀ H ₁₆	56			828	524	374	324	226	182
(6) (β -pinene)	C ₁₀ H ₁₆	55 ^b			828	524	374	324	226	182
(7) (norborneneone)	C ₇ H ₈ O	34			552	352	256	216	152	136
(8) ((M)- σ -[4]-helicene)	C ₈ H ₁₂	44			690	438	315	270	189	159
(9) (enone precursor)	C ₁₃ H ₁₈ O	71			1058	672	484	414	290	246
(10) (β -methyl-glucose)	C ₇ H ₁₄ O ₆	58			920	586	425	360	253	223
(11) (Ac-(alanine) ₂ -NH ₂)	C ₈ H ₁₅ N ₃ O ₃	59 ^b			989	630	457	387	272	240

^aNumber of normal modes between 100 and 1900 cm⁻¹. ^bNumber of normal modes between 107 and 1900 cm⁻¹.

the number of wrong signs with respect to those obtained at the aug-cc-pVQZ//aug-cc-pVQZ level of theory for molecules 1–3 are given in Table 2, respectively, for various basis set combinations. Deviations of various basis sets used to compute the Raman and ROA tensors using geometries and force fields computed at the aug-cc-pVQZ basis set are presented at the top of the table. The importance of diffuse functions in the basis set used to compute the Raman/ROA tensors is evident. The %NRMS deviation for Raman/ROA tensors computed with the cc-pVQZ basis set is roughly 3–4% for the Raman intensities and 5–7% for the ROA intensity differences, although there are no wrong signs (within 10% of the range of the reference). These results show that even the large quadruple- ζ basis set still benefits from diffuse functions. The %NRMS deviations for the aug-cc-pVTZ basis set are 0.5% or less for both Raman intensities and ROA intensity difference, and there are also no wrong signs. Removing diffuse functions from this triple- ζ basis set (cc-pVTZ) increases the %NRMS deviation to 5–7% for Raman intensities and 10–12% for ROA intensity differences. Also, there is one wrong sign for 1 and 3 and two wrong signs for 2. This effect is even more dramatic for the double- ζ basis set. The %NRMS deviations for the aug-cc-pVDZ basis set range from less than 1% for Raman intensities and 1–2% for ROA intensity differences. Removing the diffuse functions from this basis set (cc-pVDZ) increases the %NRMS deviation to 7–13% for Raman intensities and to 19–25% for ROA intensity differences and also increases the number of wrong signs. There are two wrong signs for 1, four wrong signs for 2, and three wrong signs for 3. The aug(sp)-cc-pVDZ basis set provides %NRMS deviations for Raman intensities and ROA intensity differences which are close to those obtained using full aug-cc-pVDZ basis set and introduces no wrong signs for these three molecules. The effect of removing the d type diffuse functions increases the Raman intensity %NRMS deviation by roughly 1% while increasing the ROA intensity difference %NRMS by 1–2%. Overall, diffuse d functions have a relatively small influence on the Raman/ROA tensor invariants for molecules 1–3 even though 2 contains a sulfur atom and 3 contains a chlorine atom. The small rDPS basis set provides Raman intensity and ROA intensity difference %NRMS deviations which are slightly larger than those of aug(sp)-cc-pVDZ. We note that the rDPS ROA intensity difference %NRMS deviations are slightly less (molecules 1 and 2) or the same (molecule 3) as compared to cc-pVQZ. As has previously been shown,^{42,43} and is evident from Table 1, 6-31G* is a very

poor basis set for computing Raman and ROA tensors. The %NRMS deviations range from 8 to 20% for Raman intensities and 23 to 32% for ROA intensity differences with 2–3 wrong signs introduced.

Although the objective of this study is to examine the basis set convergence of Raman intensities and ROA intensity differences, it is useful to give these %NRMS values some perspective by comparing to the experimental Raman and ROA spectra of 1. In this case, the experimental Raman and ROA spectra of 1⁸⁰ were scaled, with respect to the spectra obtained at the aug-cc-pVQZ//aug-cc-pVQZ level of theory, by matching the maximum intensity or maximum intensity difference, respectively.⁸¹ The %NRMS deviations, with respect to the experimental spectra, of the Raman intensities and ROA intensity differences obtained at the aug-cc-pVQZ//aug-cc-pVQZ level of theory are roughly 12% and 6%, respectively, and the %NRMS deviation for the frequencies is 2%.

The Raman and ROA spectra for 1 are shown in Figure 2. In this figure, Raman and ROA tensors were computed at each of the aug-cc-pVQZ, aug-cc-pVTZ, aug-cc-pVDZ, aug(sp)-cc-pVDZ, and rDPS basis sets using the aug-cc-pVQZ optimized geometry and force field. The Raman and ROA spectra obtained from the first four basis sets are almost visually indistinguishable. The Raman intensities and ROA intensity differences obtained using the rDPS basis set are overall very similar but show slight intensity increases around 400 cm⁻¹ and 950 cm⁻¹ in the ROA spectrum.

Raman intensity and ROA intensity difference %NRMS deviations obtained using the reference aug-cc-pVQZ basis set for the Raman and ROA tensors and the aug-cc-pVTZ, cc-pVTZ, 6-31G*, and aug-cc-pVDZ basis sets for the force field are also shown for molecules 1–3 in Table 2. Force fields computed at the aug-cc-pVTZ basis provide %NRMS deviations for Raman intensities and ROA intensity differences which differ from aug-cc-pVQZ by 1% or less. Removing the diffuse functions from this triple- ζ basis set (cc-pVTZ) results in %NRMS deviations of Raman intensities and ROA intensity differences which are essentially the same, with the exception of 1 where the ROA %NRMS increases by 2%. Force fields computed using the 6-31G* basis set produce %NRMS deviations of 2–4% for Raman intensities and deviations of 4–6% for ROA intensity differences. There is also one wrong sign introduced for each molecule. Force fields computed using the aug-cc-pVDZ basis set produce %NRMS deviations which are slightly higher than for 6-31G*. There is one wrong sign for 1 and 2, but no wrong signs for 3.

Table 2. %RMS Deviations of Frequencies, Raman Intensities, and ROA Intensity Differences of 1–4 for Various Basis Set Combinations^a

basis set		(1)methyloxirane			(2)methylthirane			(3)epichlorhydrin			(4)CHFCIBr						
		%NRMS deviation			NWS	%NRMS deviation			NWS	%NRMS deviation			NWS				
force field	Raman/ROA	ν	I^R+I^L	I^R-I^L		ν	I^R+I^L	I^R-I^L		ν	I^R+I^L	I^R-I^L		ν	I^R+I^L	I^R-I^L	
aug-cc-pVQZ:		0.0				0.0				0.0				0.0			
	aug-cc-pVQZ	0.0	0.0	0		0.0	0.0	0		0.0	0.0	0		0.0	0.0	0	
	aug-cc-pVTZ	0.2	0.3	0		0.2	0.5	0		0.2	0.4	0		0.6	1.6	0	
	aug-cc-pVDZ	0.8	1.2	0		0.8	2.0	0		0.7	1.4	0		2.7	6.3	0	
	aug(sp)-cc-pVDZ	1.7	3.5	0		1.6	3.3	0		1.9	3.0	0		13.4	19.4	0	
	rDPS	3.2	5.2	0		3.2	4.9	0		4.4	5.5	0		31.7	32.6	1	
	cc-pVQZ	2.8	6.0	0		3.2	7.3	0		4.3	5.2	0		8.0	10.6	0	
	cc-pVTZ	5.5	11.7	1		4.5	12.2	2		7.2	10.3	1		14.2	17.4	0	
	cc-pVDZ	13.1	24.8	2		6.9	20.7	4		11.3	19.0	3		24.7	26.0	0	
	6-31G*	19.5	31.9	2		7.6	22.6	2		11.3	25.1	3		18.7	23.0	1	
aug-cc-pVTZ:		0.1				0.1				0.2				0.2			
	aug-cc-pVQZ	0.2	1.0	0		1.0	0.6	0		1.2	0.8	0		0.7	0.7	0	
	aug-cc-pVTZ	0.3	1.1	0		1.1	0.8	0		1.3	1.0	0		1.0	2.1	0	
	aug-cc-pVDZ	0.8	1.6	0		1.3	2.1	0		1.3	1.7	0		3.0	6.9	0	
	aug(sp)-cc-pVDZ	1.7	3.8	0		1.3	3.2	0		1.3	3.6	0		13.8	19.9	0	
	rDPS	3.2	5.5	0		3.7	4.7	0		5.2	6.1	0		32.3	33.0	2	
	6-31G*	19.6	31.3	3		7.6	22.2	2		11.3	25.3	3		19.1	23.1	1	
cc-pVTZ:		0.1				0.1				0.1				0.7			
	aug-cc-pVQZ	0.8	3.0	0		0.8	0.8	0		1.3	1.1	0		1.3	1.6	0	
	aug-cc-pVTZ	0.8	3.1	0		0.9	1.0	0		1.3	1.3	0		1.6	3.1	0	
	aug-cc-pVDZ	0.8	3.5	0		1.1	2.2	0		1.3	1.9	0		3.4	8.0	0	
	aug(sp)-cc-pVDZ	1.6	5.2	1		1.4	3.2	0		1.3	3.8	0		14.3	21.7	0	
	rDPS	2.9	6.7	1		3.5	4.8	0		5.0	6.2	0		33.0	35.4	2	
	cc-pVTZ	5.9	10.9	2		4.2	12.0	3		6.8	10.4	0		14.9	19.3	0	
	cc-pVDZ	13.5	23.8	2		6.7	20.3	3		11.2	18.9	1		25.6	28.0	0	
	6-311++G(2d,2p)	3.1	8.2	0		1.6	5.7	1		2.8	6.2	0		6.1	12.5	0	
	6-311++G**	6.5	17.1	2		2.3	11.4	2		4.9	13.6	2		17.6	23.2	0	
	TZVP	9.1	22.3	2		4.4	18.8	4		7.3	16.6	1		28.0	34.7	0	
	6-31G*	19.9	30.6	3		7.7	22.0	2		11.2	25.0	3		19.4	24.8	1	
6-31G*:		1.2				1.2				1.0				2.2			
	aug-cc-pVQZ	1.9	4.8	1		2.8	5.7	1		3.5	4.1	1		2.0	4.3	0	
	aug-cc-pVTZ	1.8	5.0	1		2.9	5.8	1		3.5	4.1	1		2.2	5.8	0	
	aug-cc-pVDZ	1.8	5.4	1		3.2	5.9	1		3.2	4.3	1		3.6	10.7	0	
	aug(sp)-cc-pVDZ	2.1	7.0	1		3.4	5.4	1		3.2	5.8	1		14.0	24.7	0	
	rDPS	3.3	8.8	1		4.9	5.9	1		5.5	8.4	1		32.3	39.5	3	
	6-31G*	20.2	34.4	2		8.7	19.0	2		11.6	26.2	3		18.7	28.0	2	
aug-cc-pVDZ:		1.2				1.1				1.0				1.3			
	aug-cc-pVQZ	2.8	7.5	1		5.5	6.7	1		3.8	5.9	0		1.1	1.6	0	
	aug-cc-pVTZ	2.8	7.5	2		5.7	6.8	1		3.8	6.0	0		1.1	2.2	0	
	aug-cc-pVDZ	2.8	7.6	2		6.0	6.9	1		3.9	6.0	0		2.7	5.7	0	
	aug(sp)-cc-pVDZ	3.6	9.6	2		5.9	6.5	1		3.0	7.3	1		13.4	17.6	0	
aug(sp)-cc-pVDZ:		1.4				1.4				1.2				1.0			
	aug-cc-pVQZ	2.8	6.6	0		6.0	8.0	1		3.8	6.3	1		1.4	1.3	0	
	aug-cc-pVDZ	2.7	6.7	0		6.4	8.1	1		4.0	6.6	1		3.3	5.9	0	
	aug(sp)-cc-pVDZ	3.4	8.7	0		6.4	8.0	2		3.2	7.9	1		14.3	17.9	0	

^a %NRMS deviations with respect to the aug-cc-pVQZ basis set. NWS is the number of wrong signs within 10% of the range.

Raman and ROA %NRMS deviations obtained using the aug-(sp)-cc-pVDZ for the geometry and force field are similar to

those obtained using the aug-cc-pVDZ basis set for 1, 2, and 3. There are also no wrong signs.

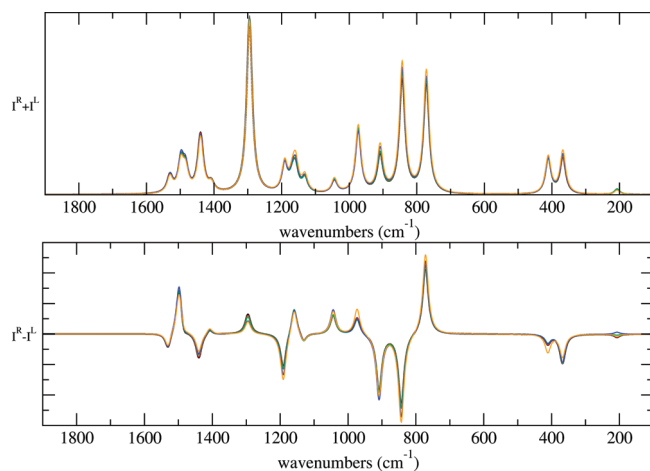


Figure 2. Raman (top) and ROA (bottom) spectra of **1** ((S)-methyl-oxirane). Geometry and force field computed using aug-cc-pVQZ. ROA tensors computed using aug-cc-pVQZ (black), aug-cc-pVTZ (red), aug-cc-pVDZ (green), aug(sp)-cc-pVDZ (blue), and rDPS (gold).

The Raman and ROA spectra for **1** for Raman and ROA tensors computed using the aug-cc-pVQZ basis set and geometry and force fields computed using the aug-cc-pVQZ, aug-cc-pVTZ, and cc-pVTZ basis sets are visually indistinguishable and are therefore not shown. The Raman and ROA spectra for **1** for Raman and ROA tensors computed using the aug-cc-pVQZ basis set and geometry and force fields computed using the cc-pVTZ, 6-31G*, and aug-cc-pVDZ basis sets are shown in Figure 3. There are apparent differences. The positions of the frequencies and the relative intensities differ, relative to cc-pVTZ, for 6-31G* and aug-cc-pVDZ starting above 700 cm^{-1} . The 6-31G* basis set provides frequencies which are slightly blue-shifted, relative to cc-pVTZ (and hence aug-cc-pVQZ), while the aug-cc-pVDZ basis set provides frequencies which are slightly red-shifted. Given the similarity of the spectra (**1**) and the small %NRMS deviations between aug-cc-pVTZ and cc-pVTZ (**1–3**), diffuse functions appear to have a minimal effect on the geometry and force field for these three molecules.

Also given in Table 2 are the Raman intensity and ROA intensity difference %NRMS deviations for other basis sets used to compute the Raman and ROA tensors using the cc-pVTZ geometry and force field. Note that the Raman and ROA %NRMS deviations are higher for the 6-311++G(2d,2p) basis set, as compared to aug-cc-pVDZ, despite the fact that the former basis set is slightly larger. As pointed out by Wiberg et al.⁸² in the context of specific rotations, the Pople style basis sets do not have a diffuse p function on hydrogen, while aug-cc-pVDZ and rDPS do. This appears to be important for Raman and ROA tensors as well. Removing the second set of polarization functions (6-311++G**) causes a dramatic increase in the ROA %NRMS deviation (to 17%, 11%, and 14%, respectively, for molecules **1–3**). The extra set of polarization functions appears to partially compensate for the lack of a p type diffuse function on hydrogen. The %NRMS deviations for Raman and ROA tensors obtained using the TZVP basis are also given in Table 2. The TZVP basis set, which lacks diffuse functions, is larger than rDPS and smaller than aug(sp)-cc-pVDZ but gives %NRMS deviations which are only slightly better than those obtained using 6-31G*.

4.2. Molecule 4. The %NRMS deviations for the frequencies, Raman intensities, and ROA intensity differences and number of

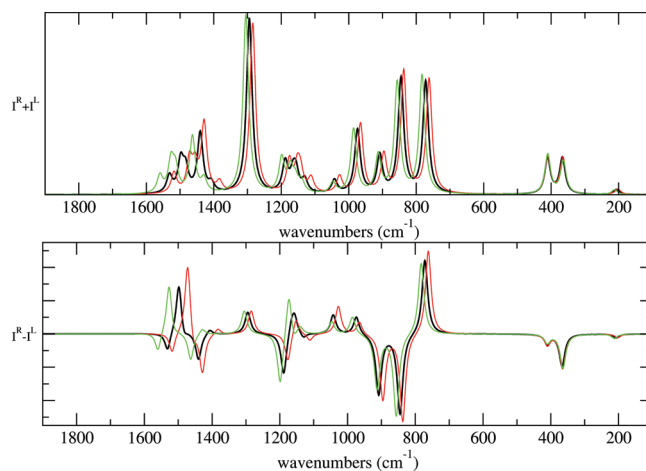


Figure 3. Raman (top) and ROA (bottom) spectra of **1** ((S)-methyl-oxirane). ROA tensors computed using aug-cc-pVQZ. Geometry and force field computed using cc-pVTZ (black), aug-cc-pVDZ (red), and 6-31G* (green).

wrong signs with respect to those obtained at the aug-cc-pVQZ//aug-cc-pVQZ level of theory for molecule **4** are given in Table 2 for various basis set combinations. The importance of diffuse functions in the basis set used to compute the Raman/ROA tensors is even more pronounced than for molecules **1–3**. The %NRMS deviation for Raman/ROA tensors computed with the cc-pVQZ basis set is roughly 8% for the Raman intensities and nearly 11% for the ROA intensity differences. These values increase to roughly 25% for the cc-pVDZ basis set. The aug(sp)-cc-pVDZ basis gives a %NRMS deviation for Raman intensities of 13% and ROA intensity differences of nearly 20% compared to 3% and 6%, respectively, for the aug-cc-pVDZ basis set. Clearly diffuse s, p, and d functions are important for **4**. The rDPS basis set, as could be expected, does not perform well for this molecule since there is only one hydrogen and only s and p diffuse functions on F, Cl, and Br. This basis set introduces one wrong sign and gives larger %NRMS deviations than those obtained using the 6-31G* basis set, for both Raman intensities and ROA intensity differences.

Raman intensity and ROA intensity difference %NRMS deviations obtained using the reference aug-cc-pVQZ basis set for the Raman and ROA tensors and the aug-cc-pVTZ, cc-pVTZ, 6-31G*, and aug-cc-pVDZ basis sets for the force field are also shown in Table 2. As was observed for molecules **1–3**, the Raman intensities and ROA intensity differences show a relatively small dependence on the presence of diffuse functions in the basis set used to compute the geometry and force field.

4.3. Molecules 5–11. The %NRMS deviations for the frequencies, Raman intensities and ROA intensity differences, and number of wrong signs with respect to those obtained at the aug-cc-pVTZ//aug-cc-pVTZ level of theory are given in Tables 3 and 4 for molecules **5–8** and **9–11**, respectively, for various basis set combinations. The results for molecules **1–4** suggest that this is reasonable given that the Raman and ROA %NRMS deviations obtained from the aug-cc-pVTZ basis set are very similar to those obtained using the much larger quadruple- ζ basis set aug-cc-pVQZ.

Deviations of various basis sets used to compute the Raman and ROA tensors using geometries and force fields computed using the aug-cc-pVTZ basis set are presented at the top of Tables 3 and 4. Again, the importance of diffuse functions for the

Table 3. %NRMS Deviations of Frequencies, Raman Intensities, and ROA Intensity Differences of Molecules 5–8 for Various Basis Set Combinations^a

basis set		(5) α -pinene			(6) β -pinene			(7)norborneneone			(8)(M)- σ -[4]-helicene						
		%NRMS deviation			%NRMS deviation			%NRMS deviation			%NRMS deviation						
force field	Raman/ROA	ν	I^R+I^L	I^R-I^L	NWS	ν	I^R+I^L	I^R-I^L	NWS	ν	I^R+I^L	I^R-I^L	NWS	ν	I^R+I^L	I^R-I^L	NWS
aug-cc-pVTZ:		0.0			0.0			0.0			0.0						
	aug-cc-pVTZ	0.0	0.0	0		0.0	0.0	0		0.0	0.0	0		0.0	0.0	0	
	aug-cc-pVDZ	0.3	0.5	0		0.2	0.5	0		0.4	0.4	0		0.2	0.4	0	
	aug(sp)-cc-pVDZ	0.5	1.0	0		0.4	0.8	0		1.8	1.3	0		0.4	1.0	0	
	rDPS	1.1	1.8	0		1.9	1.6	0		3.6	2.3	0		1.2	2.7	1	
cc-pVTZ:		0.1			0.1			0.1			0.1						
	aug-cc-pVTZ	0.1	0.5	0		0.3	0.6	0		0.6	0.4	0		0.2	0.3	0	
	aug-cc-pVDZ	0.3	0.7	0		0.3	0.7	0		0.8	0.7	0		0.2	0.5	0	
	aug(sp)-cc-pVDZ	0.5	1.1	0		0.6	1.0	0		1.9	1.4	0		0.4	1.1	0	
	rDPS	1.1	1.9	0		1.9	1.8	0		3.7	2.4	0		1.2	2.7	0	
	cc-pVTZ	3.2	6.5	0		5.6	7.8	1		7.8	4.9	0		3.6	7.2	1	
	cc-pVDZ	6.7	15.0	3		8.5	15.4	3		11.8	9.5	1		6.6	15.4	1	
	6-311++G(2d,2p)	0.6	1.9	0		0.6	1.7	0		1.2	1.0	0		0.4	1.3	0	
	6-311++G**	1.7	4.9	0		1.1	4.6	0		2.8	2.3	0		1.1	2.9	0	
	TZVP	3.4	9.1	1		4.5	9.7	2		6.6	5.4	0		3.4	6.3	1	
	6-31G*	10.3	22.3	10		10.1	20.4	6		13.6	12.1	1		8.3	21.5	1	
6-31G*:		1.0			1.0			0.8			0.9						
	aug-cc-pVTZ	1.1	4.4	1		1.8	4.9	2		3.1	1.6	0		1.7	2.3	0	
	aug-cc-pVDZ	1.1	4.3	1		1.8	4.8	2		3.2	1.8	0		1.8	2.3	0	
	aug(sp)-cc-pVDZ	1.0	4.9	1		2.0	5.0	2		3.9	2.6	0		1.8	2.6	0	
	rDPS	1.4	5.0	1		2.9	5.1	2		5.1	3.1	0		2.4	3.6	0	
	6-31G*	10.2	28.7	9		10.3	20.1	7		14.1	12.9	1		9.0	22.1	4	
aug-cc-pVDZ:		0.6			0.6			0.6			0.7						
	aug-cc-pVTZ	4.9	10.3	2		2.3	4.4	1		5.7	3.1	0		1.6	3.4	1	
	aug-cc-pVDZ	4.9	10.2	2		2.2	4.3	1		5.7	3.3	0		1.7	3.5	1	
	aug(sp)-cc-pVDZ	4.8	10.5	2		2.3	4.4	1		6.0	3.6	0		1.6	3.7	1	
aug(sp)-cc-pVDZ:		0.9			0.9			0.7			1.3						
	aug-cc-pVDZ	3.3	8.8	3		2.5	4.6	3		3.1	3.4	0		2.6	3.6	0	
	aug(sp)-cc-pVDZ	3.2	8.8	2		2.4	4.9	3		3.8	4.0	0		2.2	3.7	0	

^a %NRMS deviations with respect to the aug-cc-pVTZ basis set. NWS is the number of wrong signs within 10% of the range.

Raman/ROA tensors is evident. The %NRMS deviations for Raman intensities and ROA intensity differences for molecules 5–11 show trends similar to those for 1–4. Raman and ROA tensors computed using the aug-cc-pVDZ basis set provide %NRMS deviations for both Raman intensities and ROA intensity differences that differ from those obtained using the aug-cc-pVTZ basis set by less than 1%. There are also no wrong signs. The aug(sp)-cc-pVDZ basis set provides %NRMS deviations for Raman intensities and ROA intensity differences which are close to those obtained using full aug-cc-pVDZ basis set and introduces no wrong signs for these seven molecules. The effect of removing the d type diffuse functions increases both the Raman intensity and ROA intensity difference %NRMS deviations by less than 2%. With the exception of **10**, the small rDPS basis set provides Raman intensity and ROA intensity difference %NRMS deviations which are a few percent larger than those of aug(sp)-cc-pVDZ.

The Raman and ROA spectra for **5** are shown in Figure 4, and the ROA spectra of **9** are shown at the top of Figure 6. In these figures, the Raman and ROA tensors were computed at each of the aug-cc-pVTZ, aug-cc-pVDZ, aug(sp)-cc-pVDZ, and rDPS

basis sets using the aug-cc-pVTZ optimized geometry and force field. The Raman and ROA spectra obtained from these basis sets are almost visually indistinguishable.

Raman intensity and ROA intensity difference %NRMS deviations obtained using the reference aug-cc-pVTZ basis set for the Raman and ROA tensors and each of the cc-pVTZ, 6-31G*, and aug-cc-pVDZ basis sets for the force field are also given in Tables 3 and 4 for molecules 5–8 and 9–11, respectively. For molecules 5–9, force fields computed using the cc-pVTZ basis set provide %NRMS deviations for Raman intensities and ROA intensity differences which differ from aug-cc-pVTZ by 1% or less. These values are higher for molecules **10** and **11**. For molecules 5–9, force fields computed using the 6-31G* basis set produce %NRMS deviations of 1–3% for Raman intensities and deviations of 2–5% for ROA intensity differences. Again, these values are higher for molecules **10** and **11**, where the Raman intensity %NRMS deviations are 8% and 7%, respectively, and the ROA intensity difference %NRMS deviations are 11% and 7%, respectively. With the exception of molecules **10** and **11**, force fields computed using either the aug-cc-pVDZ or aug(sp)-cc-pVDZ basis

Table 4. %NRMS Deviations of Frequencies, Raman Intensities, and ROA Intensity Differences of Molecules 9–11 for Various Basis Set Combinations^a

basis set		(9)enone			(10)methyl- β -D-glucose (gg)			(11)Ac-(alanine) ₂ -NH ₂					
		%NRMS deviation			NWS	%NRMS deviation			NWS	%NRMS deviation			NWS
force field	Raman/ROA	ν	I ^R +I ^L	I ^R -I ^L		ν	I ^R +I ^L	I ^R -I ^L		ν	I ^R +I ^L	I ^R -I ^L	
aug-cc-pVTZ:		0.0				0.0				0.0			
	aug-cc-pVTZ	0.0	0.0	0		0.0	0.0	0		0.0	0.0	0	0
	aug-cc-pVDZ	0.1	0.5	0		0.3	0.7	0		0.6	0.5	0	0
	aug(sp)-cc-pVDZ	0.1	1.2	0		1.6	2.5	0		1.9	1.4	0	0
	rDPS	1.0	2.4	0		6.0	8.6	1		3.7	4.4	0	0
cc-pVTZ:		0.1				0.2				0.2			
	aug-cc-pVTZ	0.6	0.5	0		1.5	2.9	0		2.9	3.6	0	0
	aug-cc-pVDZ	0.6	0.7	0		1.6	2.9	0		3.1	3.7	0	0
	aug(sp)-cc-pVDZ	0.6	1.1	0		2.5	3.4	0		3.5	4.6	0	0
	rDPS	1.4	2.3	0		6.3	8.5	2		4.8	7.3	0	0
	cc-pVTZ	2.7	5.4	0		6.9	11.6	4		8.4	7.4	0	0
	cc-pVDZ	4.1	15.0	4		12.9	19.3	10		14.5	13.8	4	4
	6-311++G(2d,2p)	0.7	1.2	0		2.5	4.4	0		3.4	4.1	0	0
	6-311++G**	0.6	3.6	0		3.8	6.9	0		4.5	5.9	0	0
	TZVP	2.0	7.1	0		7.6	14.1	4		8.8	8.4	1	1
	6-31G*	4.2	24.3	8		15.9	25.0	10		19.0	17.3	6	6
6-31G*:		0.9				1.3				1.1			
	aug-cc-pVTZ	1.4	3.4	1		7.8	10.9	1		6.8	7.1	1	1
	aug-cc-pVDZ	1.4	3.4	1		8.2	11.0	1		7.0	7.1	1	1
	aug(sp)-cc-pVDZ	1.4	3.4	1		8.3	12.6	1		7.4	8.1	2	2
	rDPS	1.8	3.9	1		10.4	19.1	3		8.2	10.5	2	2
	6-31G*	4.3	25.2	7		17.6	37.1	9		19.2	18.7	8	8
aug-cc-pVDZ:		0.5				1.5				0.7			
	aug-cc-pVTZ	1.4	6.6	4		3.2	3.3	0		6.5	4.5	1	1
	aug-cc-pVDZ	1.4	6.6	4		3.2	3.5	0		6.6	4.6	1	1
	aug(sp)-cc-pVDZ	1.4	7.0	4		3.6	4.7	0		7.0	4.9	1	1
aug(sp)-cc-pVDZ:		0.8				0.8				0.9			
	aug-cc-pVDZ	1.7	7.4	1		3.3	3.7	0		6.0	4.6	0	0
	aug(sp)-cc-pVDZ	1.7	7.8	2		3.7	4.8	0		6.4	5.0	1	1

^a %NRMS deviations with respect to the aug-cc-pVTZ basis set. NWS is the number of wrong signs within 10% of the range.

sets produce %NRMS deviations which are the same or higher than those for 6-31G*. These results indicate that molecules **10** and **11** are more sensitive to the presence of diffuse functions in the basis set used to compute the force field.

The dependence of the Raman (for molecule **5**) and ROA (for molecules **5** and **9**) spectra on the basis set used for the geometry and force field are shown in Figure 5 and at the bottom of Figure 6, respectively. As can be observed in these figures, the Raman and ROA spectra computed using the aug-cc-pVTZ and cc-pVTZ basis sets are visually indistinguishable. The positions of the frequencies and the relative intensities differ, relative to aug-cc-pVTZ (or cc-pVTZ), for 6-31G* and aug-cc-pVDZ starting above 700 cm⁻¹. The 6-31G* basis set provides frequencies which are slightly blue-shifted, while the aug-cc-pVDZ basis set provides frequencies which are slightly red-shifted.

Also given in Tables 3 (molecules **5**–**8**) and 4 (molecules **9**–**11**) are the Raman intensity and ROA intensity difference %NRMS deviations for other basis sets used to compute the Raman and ROA tensors using the cc-pVTZ geometry and force field. The same trends can be observed as for molecules **1**–**3**.

The Sadlej basis set^{74,75} has been shown to yield Raman intensities⁴² and ROA intensity differences^{43,44} of a quality close to aug-cc-pVTZ. Raman intensity and ROA intensity difference %NRMS deviations for the Sadlej basis set were found to be equivalent (within 1%) to the aug-cc-pVTZ deviations and are therefore not shown. The large LPol-ds basis set of Sadlej⁷⁶ has been shown to correctly predict the sign and magnitude of the optical rotation for **6** (β -pinene) at long wavelengths and to significantly improve the short wavelength prediction, as compared to aug-cc-pVDZ and aug-cc-pVTZ. However, Raman and ROA %NRMS deviations computed using the LPol-ds basis set were found to be essentially equivalent to those obtained using the aug-cc-pVTZ basis set and are likewise not shown. Raman intensity and ROA intensity difference %NRMS deviations obtained using the (doubly augmented) d-aug-cc-pVTZ basis set were found to be the same as those obtained with the aug-cc-pVTZ basis set. The (doubly augmented) d-aug-cc-pVDZ basis set gave Raman intensity and ROA intensity difference %NRMS deviations which were either slightly closer to those obtained with aug-cc-pVTZ or the same as aug-cc-pVDZ.

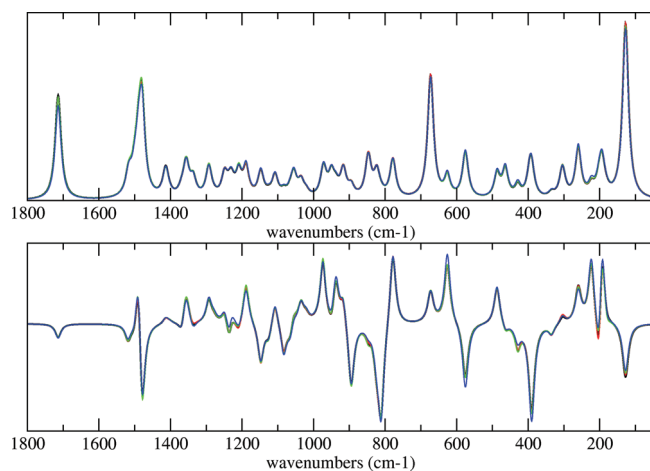


Figure 4. Raman (top) and ROA (bottom) spectra of **5** ((1S,5S)- α -pinene). Geometry and force field computed using aug-cc-pVTZ. ROA tensors computed using aug-cc-pVTZ (black), aug-cc-pVDZ (red), aug(sp)-cc-pVDZ (green), and rDPS (blue).

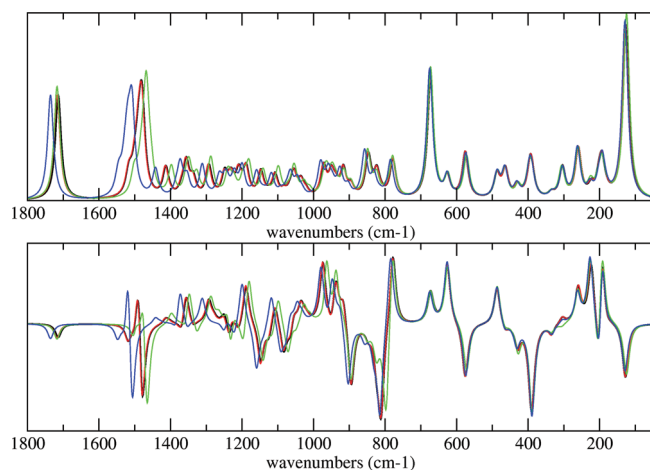


Figure 5. Raman (top) and ROA (bottom) spectra of **5** ((1S,5S)- α -pinene). ROA tensors computed using aug-cc-pVTZ. Geometry and force field computed using aug-cc-pVTZ (black), cc-pVTZ (red), aug-cc-pVDZ (green), and 6-31G* (blue).

4.4. Timing Differences between the “One-Step” and “Two-Step” Procedures. Relative timings for the various basis combinations are given in Table 5 for molecule **9**, the largest molecule examined here. Timings in this table assume 100 CPU units for the aug-cc-pVTZ//aug-cc-pVTZ level of theory. Data are given with and without accounting for the geometry optimization time, which assumes nine steps. Compared to the aug-cc-pVTZ//aug-cc-pVTZ level of theory, the aug-cc-pVDZ//cc-pVTZ combination reduces the amount of CPU time by more than a factor of 10. The aug-cc-pVDZ//6-31G* combination reduces time by roughly factor of 20, while the aug(sp)-cc-pVDZ//6-31G* combination reduces the time by more than a factor of 30. When the geometry optimization step is included, the aug(sp)-cc-pVDZ//6-31G* combination reduces the CPU time by a factor of 40, relative to aug-cc-pVTZ//aug-cc-pVTZ. The aug(sp)-cc-pVDZ basis set provides nearly a factor of 2 reduction in cost over the aug-cc-pVDZ basis set for computing just the Raman/ROA tensors via the “two-step” procedure. The time to

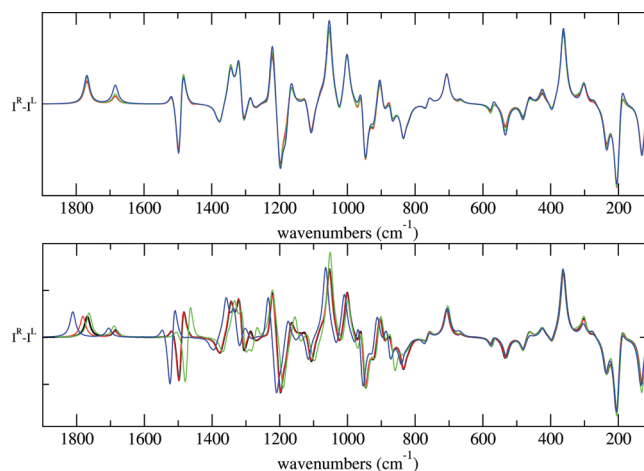


Figure 6. ROA spectra of **9**. Top: Geometry and force field computed using aug-cc-pVTZ. ROA tensors computed using aug-cc-pVTZ (black), aug-cc-pVDZ (red), aug(sp)-cc-pVDZ (green), and rDPS (blue). Bottom: ROA tensors computed using aug-cc-pVTZ. Geometry and force field computed using aug-cc-pVTZ (black), cc-pVTZ (red), aug-cc-pVDZ (green), and 6-31G* (blue).

Table 5. Relative Timings for Various Basis Set Combinations for Molecule 9^a

basis set		time	time (including geom. opt) ^b
force field	Raman/ROA		
aug-cc-pVTZ:			
	aug-cc-pVTZ ^c	100.0	100.0
cc-pVTZ:			
	aug-cc-pVDZ	9.5	9.8
	aug(sp)-cc-pVDZ	7.1	7.7
	rDPS	5.3	6.3
6-31G*:			
	aug-cc-pVDZ	5.4	4.6
	aug(sp)-cc-pVDZ	3.0	2.5
	rDPS	1.2	1.1
aug-cc-pVDZ:			
	aug-cc-pVDZ ^c	5.5	5.8
	aug-cc-pVDZ	7.8	7.7
aug(sp)-cc-pVDZ:			
	aug(sp)-cc-pVDZ ^c	3.1	3.3
	aug(sp)-cc-pVDZ	4.2	4.2

^aTime to compute just the Raman/ROA tensors using the “two-step” procedure: aug-cc-pVDZ = 5.1; aug(sp)-cc-pVDZ = 2.7; rDPS = 0.9. ^bAssuming nine optimization steps. ^cComputed using the “one-step” procedure.

compute both Raman intensities and ROA intensity differences using the aug-cc-pVDZ//6-31G* combination is essentially the same as the aug-cc-pVDZ//aug-cc-pVDZ combination (obtained via the “one-step” procedure), not including the geometry optimization step. However, when the geometry optimization step is included, the aug-cc-pVDZ//6-31G* combination shows a slight advantage. This is also the case with aug(sp)-cc-pVDZ//6-31G* and aug(sp)-cc-pVDZ//aug(sp)-cc-pVDZ. This difference will increase with the size of the molecule, and also if more geometry

Table 6. %NRMS Deviations of ROA Intensity Differences with and without the Electric-Dipole–Electric-Quadrupole Contribution^a

molecule	ref		aug(sp)-cc-pVDZ//6-31G*		
	$I^R - I^L$	$I^R - I^L$	NWS	$I^R - I^L$	NWS
	(without EQ)	(including EQ)	(without EQ)	(without EQ)	(without EQ)
(1) (methyloxirane)	3.1	7.0	1	5.9	1
(2) (methylthirane)	2.1	8.1	1	7.2	1
(3) (epichlorhydrin)	2.7	5.9	1	4.8	1
(4) (CHFCIBr)	2.0	24.8	0	25.8	1
(5) (α -pinene)	2.9	4.9	1	4.4	1
(6) (β -pinene)	3.2	5.0	2	5.5	2
(7) (norborneneone)	1.7	2.6	0	2.0	0
(8) ((M)- σ -[4]-helicene)	3.2	2.6	0	4.2	0
(9) (enone precursor)	3.7	3.4	1	5.1	0
(10) (β -methyl-glucose)	3.8	13.5	3	12.0	3
(11) (Ac-(alanine) ₂ -NH ₂)	3.0	8.1	2	7.9	2

^a aug-cc-pVQZ//aug-cc-pVQZ for molecules 1–4. aug-cc-pVTZ//aug-cc-pVTZ for molecules 5–11.

optimization steps are required (i.e., floppy molecules). The aug(sp)-cc-pVDZ//6-31G* combination provides roughly a factor of 2 reduction in cost over the aug-cc-pVDZ//6-31G* combination, including the geometry optimization step. The rDPS/6-31G* combination provides roughly a factor of 4 reduction over aug-cc-pVDZ//6-31G*, including the geometry optimization steps. The aug(sp)-cc-pVDZ//6-31G* and rDPS//6-31G* combinations would become even more advantageous for larger molecules.

4.5. Electric-Dipole–Electric-Quadrupole Contribution. Lubert et al.⁴⁶ examined the importance of the electric-dipole–electric-quadrupole contribution to ROA intensity differences ($\beta_{A_i}^2$ in eqs 1 and 2). With the exception of some C–H stretching vibrations, they found the electric-dipole–electric-quadrupole contribution to the SCP backscattering ROA intensity differences to be small. They found that neglecting the electric-dipole–electric-quadrupole tensor resulted in an overall small change to the spectra below 2000 cm⁻¹. (Molecules 1 (methyloxirane), 3 (epichlorhydrin), 6 (β -pinene), and 8 ((M)- σ -[4]-helicene) were included in their study.) Neglecting the electric-dipole–electric-quadrupole contribution reduces the number of CPKS equations from 41 to 21 for the “two-step” procedure (and from 11 to 6 for the “one-step” procedure). Although the number of CPKS equations is reduced by nearly a factor of 2, this results in only a modest speedup (roughly 1.4 for molecule 9) in computing the Raman and ROA tensors using the “two-step” procedure.

The effect of neglecting the electric-dipole–electric-quadrupole contribution on the ROA intensity difference %NRMS deviations is shown in Table 6. For molecules 1–4, excluding the electric-dipole–electric-quadrupole contribution gives a %NRMS deviation for the ROA intensity differences of 2–3%, with respect to aug-cc-pVQZ//aug-cc-pVQZ level of theory. For molecules 5–11, the %NRMS deviation is 2–4% with respect to aug-cc-pVTZ//aug-cc-pVTZ. Interestingly, the %NRMS deviation obtained from excluding the electric-quadrupole contribution becomes slightly less than the %NRMS, which includes this contribution at the aug(sp)-cc-pVDZ//6-31G* level of theory for molecules 1–3, 5, 8, 10, and 11. For molecules 6, 7, and 9, the

%NRMS deviation obtained from excluding the electric-quadrupole contribution is just slightly larger. Although not shown, this same trend is observed at the rDPS//6-31G* level of theory. On the basis of these limited results (and those of ref 46), neglecting the electric-dipole–electric-quadrupole tensor contribution appears to be a reasonable approximation for computing SCP backscattering ROA spectra, especially when applied to large molecules where the aug(sp)-cc-pVDZ//6-31G* or rDPS//6-31G* level of theory is practical. It remains to be seen if this is, in fact, general.

5. CONCLUSIONS

In this study, we have conducted a systematic investigation of the basis set dependence of the backscattering vibrational Raman intensities and Raman Optical Activity (ROA) intensity differences. We have separately quantified the basis set dependence of the Raman/ROA tensor invariants and the force field on the resulting Raman intensities and ROA intensity differences. We observe that the basis set requirements for Raman/ROA tensor invariants and the force field are not similar. Raman/ROA tensor invariants require basis sets with diffuse functions while geometry optimizations and force field calculations typically do not. Given these observations, we conclude that the “two-step” procedure (n+1 rule) is more efficient than the “one-step” procedure (2n+1 rule) for computing accurate Raman intensities and ROA intensity differences, especially for large molecules.

Raman/ROA tensor invariants computed using the aug-cc-pVTZ basis set provide Raman intensities and ROA intensity differences which are essentially equivalent to those obtained using the much larger aug-cc-pVQZ basis set (for molecules 1–4). More importantly, Raman intensities and ROA intensity differences obtained using the aug-cc-pVDZ basis set for the Raman/ROA tensor invariants are of nearly the same quality as those obtained using the aug-cc-pVTZ basis set. We find that modifying the aug-cc-pVDZ basis set by removing the set of diffuse d functions (while keeping the diffuse s and p sets) results in a basis set, denoted as aug(sp)-cc-pVDZ, which is significantly faster without much loss in the overall accuracy. In addition, we find that the popular rDPS basis set introduced by Zuber and Hug⁴³ offers a good compromise between accuracy and cost.

Geometries and force fields computed using the cc-pVTZ basis set provide Raman intensities and ROA intensity differences which are essentially equivalent to those obtained using the aug-cc-pVTZ basis, which are in turn are nearly equivalent to those obtained using the aug-cc-pVQZ basis set (for molecules 1–4). At least for the types of molecules examined here, geometries and force fields computed at the aug-cc-pVTZ basis set offer no additional advantage over cc-pVTZ. Geometries and force fields computed using the 6-31G* basis set provide Raman intensities and ROA intensity differences which are less accurate, but still acceptable. With the exception of 10 and 11, the 6-31G* basis set (for the force field determination) provides Raman intensities and ROA intensity differences which are the same or better than those of aug-cc-pVDZ. The aug(sp)-cc-pVDZ basis set is slightly less accurate but still provides reasonable results. The 6-31G* basis set offers a significant computational advantage for geometry optimizations and force field calculations, especially for large molecules, where the geometry optimization step can require a significant amount of computational time (on the same order as computing the force field), especially if the molecule has multiple low energy conformations which need to be explored.

If one assumes nine optimization steps, geometry optimization takes roughly one-half of the time required to compute the frequencies for molecule 9.

For small molecules, we recommend the aug-cc-pVDZ basis for computing the Raman/ROA tensor invariants combined with geometry optimizations and force fields computed using the cc-pVTZ basis set (aug-cc-pVDZ//cc-pVTZ level of theory), for intermediate size molecules, aug(sp)-cc-pVDZ//cc-pVTZ or rDPS//6-31G*, and for large molecules, either the aug(sp)-cc-pVDZ//6-31G* or rDPS//6-31G* levels of theory. If the “one-step” procedure is the only option, then the aug-cc-pVDZ//aug-cc-pVDZ or aug(sp)-cc-pVDZ//aug(sp)-cc-pVDZ levels of theory offer the only reasonable choice. Raman/ROA tensor invariants computed using either the aug(sp)-cc-pVDZ or smaller rDPS basis sets, in combination with the geometry optimizations and force fields computed using the 6-31G* basis set, are a reasonable choice in terms of accuracy versus cost and should be applicable to the study of molecules such as peptides, proteins, carbohydrates, and natural products.

AUTHOR INFORMATION

Corresponding Author

*E-mail: cheese@gaussian.com.

ACKNOWLEDGMENT

The authors would like to thank Professor Larry Nafie (Syracuse University and Biotools) and Dr. Hank Lee (Biotools) for kindly providing the experimental Raman and ROA spectra of 1.

REFERENCES

- Barron, L. D.; Buckingham, A. D. *Mol. Phys.* **1971**, *20*, 1111.
- Barron, L.; Bogaard, M. P.; Buckingham, A. D. *J. Am. Chem. Soc.* **1973**, *95*, 603.
- Hug, W.; Hangartner, G. A. *J. Raman Spectrosc.* **1999**, *30*, 841–852.
- Barron, L.; Buckingham, A. D. *Chem. Phys. Lett.* **2010**, *492*, 199–213.
- Pecul, M. *Chem. Phys. Lett.* **2006**, *427*, 166–176.
- Jalkanen, K. J.; Degtyarenko, I. M.; Neiminen, R. M.; Cao, X.; Nafie, L. A.; Zhu, F.; Barron, L. *Theor. Chem. Acc.* **2008**, *119*, 191–210.
- Herrmann, C.; Ruud, K.; Reiher, M. *ChemPhysChem* **2006**, *7*, 2189–2196.
- Jacob, C. R.; Lubber, S.; Reiher, M. *J. Phys. Chem. B* **2009**, *113*, 6558–6573.
- Jacob, C. R.; Lubber, S.; Reiher, M. *Chem.—Eur. J.* **2009**, *15*, 13491–13508.
- Mukhopadhyay, P.; Zuber, G.; Beratan, D. N. *Biophys. J.* **2008**, *95*, 5574–5586.
- Kapitán, J.; Fujiang, Z.; Hecht, L.; Gardiner, J.; Seebach, D.; Barron, L. *Angew. Chem., Int. Ed.* **2008**, *47*.
- Luber, S.; Reiher, M. *J. Phys. Chem. B* **2010**, *114*, 1057–1063.
- Luber, S.; Reiher, M. *Chem. Phys.* **2008**, *346*, 212–223.
- Macleod, N. A.; Johannessen, C.; Hecht, L.; Barron, L.; Simons, J. P. *Int. J. Mass Spectrom.* **2006**, *253*, 193–200.
- Luber, S.; Reiher, M. *J. Phys. Chem. A* **2009**, *113*, 8268–8277.
- Kaminský, J.; Kapitán, J.; Baumruk, V.; Bednářová, L.; Bouř, P. *J. Phys. Chem. A* **2009**, *3594*–3601.
- Hug, W.; Zuber, G.; de Meijere, A.; Khlebnikov, A. F.; Hansen, H.-J. *Helv. Chim. Acta* **2001**, *84*, 1–21.
- Liégeois, V.; Champagne, B. *J. Comput. Chem.* **2008**, *30*, 1261–1278.
- Kapitán, J.; Johannessen, C.; Bouř, P.; Hecht, L.; Barron, L. *Chirality* **2009**, *21*, E4–E12.
- Fedorovsky, M.; Gerlach, H.; Hug, W. *Helv. Chim. Acta* **2009**, *92*, 1451–1465.
- Liégeois, V.; Quinet, O.; Champagne, B.; Haesler, J.; Zuber, G.; Hug, W. *Vib. Spectrosc.* **2006**, *42*, 309–316.
- Hug, W.; Haesler, J. *Int. J. Quantum Chem.* **2005**, *104*, 695–715.
- Herrmann, C.; Neugebauer, J.; Reiher, M. *New J. Chem.* **2007**, *31*, 818–831.
- Liégeois, V. *ChemPhysChem* **2009**, *10*, 2017–2025.
- Lovchik, M. A.; Frater, G.; Goeke, A.; Hug, W. *Chem. Biodivers.* **2008**, *5*, 126–139.
- Pecul, M. *Chirality* **2009**, *21*, E98–E104.
- Hecht, L.; Nafie, L. A. *Mol. Phys.* **1991**, *72*, 441–469.
- Nafie, L. A.; Che, D. In *Modern Nonlinear Optics, Part 3*; Evans, M., Kielich, S., Eds.; Wiley: New York, 1994; *Advances in Chemical Physics* Vol. 85, pp 105–149.
- Nafie, L. A. *Annu. Rev. Phys. Chem.*; Annual Reviews, Inc.: Palo Alto, CA, 1997; Vol. 48, pp 357–386.
- Nafie, L. A. In *Encyclopedia of Spectroscopy and Spectrometry*; Tranter, G., Holmes, J., Lindon, J., Eds.; Academic Press: London, 2000; pp 1976–1985.
- Barron, L. *Molecular Light Scattering and Optical Activity*, 2nd ed.; Cambridge University Press: Cambridge, U. K., 2004.
- Hug, W. In *Handbook of Vibrational Spectroscopy*; Chalmers, J. M., Griffiths, P. R., Eds.; John Wiley and Sons: Chichester, U. K., 2002; pp 745–758.
- Ruud, K.; Thorvaldsen, A. J. *Chirality* **2009**, *21*, E54–E67.
- Liégeois, V.; Ruud, K.; Champagne, B. *J. Chem. Phys.* **2007**, *127*, 204105.
- Bast, R.; Ekström, U.; Gao, B.; Helgaker, T.; Ruud, K.; Thorvaldsen, A. J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2627–2651.
- Thorvaldsen, A. J.; Ruud, K.; Kristensen, K.; Jørgensen, P.; Coriani, S. *J. Chem. Phys.* **2008**, *129*, 214108.
- Thorvaldsen, A. J.; Ruud, K.; Jaszuński, M. *J. Phys. Chem. A* **2008**, *112*, 11942–11950.
- Furche, F. *J. Chem. Phys.* **2007**, *126*, 201104.
- Reiher, M.; Neugebauer, J. *J. Chem. Phys.* **2003**, *118*, 1634–1641.
- Luber, S.; Neugebauer, J.; Reiher, M. *J. Chem. Phys.* **2009**, *130*, 064105.
- Luber, S.; Reiher, M. *ChemPhysChem* **2009**, *10*, 2049–2057.
- Halls, M. D.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 8819.
- Zuber, G.; Hug, W. *J. Phys. Chem. A* **2004**, *108*, 2108–2119.
- Reiher, M.; Liégeois, V.; Ruud, K. *J. Phys. Chem. A* **2005**, *109*, 7567–7574.
- Bose, P. K.; Polavarapu, P. L.; Barron, L. D.; Hecht, L. *J. Phys. Chem.* **1990**, *94*, 1734–1740.
- Luber, S.; Herrmann, C.; Reiher, M. *J. Phys. Chem. B* **2008**, *112*, 2218–2232.
- Costante, J.; Hecht, L.; Polavarapu, P. L.; Collet, A.; Barron, L. D. *Angew. Chem., Int. Ed.* **1997**, *36*, 885–887.
- Che, J.; Nafie, L. A. *Chem. Phys. Lett.* **1992**, *189*, 35–46.
- Ruud, K.; Helgaker, T.; Bouř, P. *J. Phys. Chem. A* **2002**, *106*, 7448–7455.
- Barron, L. D.; Hecht, L.; Gargaro, A. R.; Hug, W. *J. Raman Spectrosc.* **1990**, *21*, 375–379.
- Yu, G. S.; Freedman, T. B.; Nafie, L. A. *J. Raman Spectrosc.* **1995**, *26*, 733–743.
- Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J. *J. Phys. Chem. A* **2001**, *105*, 5356–5371.
- Ruud, K.; Stephens, P. J.; Devlin, F. J.; Taylor, P. R.; Cheeseman, J. R.; Frisch, M. J. *Chem. Phys. Lett.* **2003**, *373*, 606–614.
- Jean-Charles, K. L.; Camara, C.; Dumas, F. First Asymmetric Synthesis of Suberosenone and Suberosanone: Absolute Stereochemistry Assignment of Natural Suberosanes. Poster at Chirality 2004, the 16th International Symposium on Chirality, ISCD 16, New York.
- Stephens, P. J.; McCann, D. M.; Devlin, F. J.; Smith, I., A. B. *J. Nat. Prod.* **2006**, *69*, 1055–1064.
- Cheeseman, J. R.; Shaik, M. J.; Popelier, P. L. A.; Blanch, E. W. *J. Am. Chem. Soc.* **2011**.

- (57) Bell, A.; Barron, L.; Hecht, L. *Carbohydr. Res.* **1994**, *257*, 11–24.
- (58) Polavarapu, P. L.; Deng, Z. *Faraday Discuss.* **1994**, *99*, 151–163.
- (59) Hug, W. *Chem. Phys.* **2001**, *264*, 53–69.
- (60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian Development Version*, Revision H.09+; Gaussian, Inc.: Wallingford, CT, 2010.
- (61) London, F. *J. Phys-Paris* **1937**, *8*, 397–409.
- (62) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789–807.
- (63) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (64) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; DeFrees, D. J.; Pople, J. A.; Gordon, M. S. *J. Chem. Phys.* **1982**, *77*, 3654–3665.
- (65) Hariharan, P. C.; Pople, J. A. *Theor. Chem. Acc.* **1973**, *28*, 213–222.
- (66) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294–301.
- (67) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.
- (68) Gill, P. M. W.; Johnson, B. G.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1992**, *197*, 499–505.
- (69) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639–5648.
- (70) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.
- (71) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (72) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (73) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358–1371.
- (74) Sadlej, A. J. *Collect. Czech. Chem. C.* **1988**, *53*.
- (75) Sadlej, A. J. *Theor. Chem. Acc.* **1991**, *79*, 123.
- (76) Baranowska, A.; Łączkowski, K. Z.; Sadlej, A. J. *J. Comput. Chem.* **2009**, *31*, 1176–1181.
- (77) Schaefer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (78) Schaefer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (79) Ruud, K.; Skaane, H.; Helgaker, T.; Bak, K.; Jørgensen, P. *J. Am. Chem. Soc.* **1994**, *116*, 10135–10140.
- (80) Lee, H.; Nafie, L. Private communication.
- (81) Absolute Raman intensities and ROA intensity differences were not measured experimentally.
- (82) Wiberg, K. B.; Wang, Y.; Vaccaro, P. H.; Cheeseman, J. R.; Trucks, G. W.; Frisch, M. J. *J. Phys. Chem. A* **2004**, *108*, 32–38.

Microscopic Simulations of Charge Transport in Disordered Organic Semiconductors

Victor Rühle,[†] Alexander Lukyanov,[†] Falk May,[†] Manuel Schrader,[†] Thorsten Vehoff,[†] James Kirkpatrick,[‡] Björn Baumeier,^{†,§} and Denis Andrienko^{*,†,§,||}

[†]Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany

[‡]Oxford Centre for Collaborative Applied Mathematics (OCCAM), University of Oxford, St Giles' 24-29, OX1 3LB Oxford, United Kingdom

[§]Institute for Pure and Applied Mathematics, University of California Los Angeles, 460 Portola Plaza, Los Angeles, California 90095, United States

^{||}Center for Organic Photonics and Electronics and School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

 Supporting Information

ABSTRACT: Charge carrier dynamics in an organic semiconductor can often be described in terms of charge hopping between localized states. The hopping rates depend on electronic coupling elements, reorganization energies, and driving forces, which vary as a function of position and orientation of the molecules. The exact evaluation of these contributions in a molecular assembly is computationally prohibitive. Various, often semiempirical, approximations are employed instead. In this work, we review some of these approaches and introduce a software toolkit which implements them. The purpose of the toolkit is to simplify the workflow for charge transport simulations, provide a uniform error control for the methods and a flexible platform for their development, and eventually allow in silico prescreening of organic semiconductors for specific applications. All implemented methods are illustrated by studying charge transport in amorphous films of tris-(8-hydroxyquinoline)aluminum, a common organic semiconductor.

I. INTRODUCTION

The progress currently observed in the field of organic electronics is a result of a combined effort of several communities. Synthetic chemists have identified classes of promising compounds, ranging from small conjugated molecules to self-assembling oligomers and conjugated polymers, and developed new synthetic routes, improving both stability and processability of the materials.^{1–7} At the same time, material processing, such as doping, annealing, use of a secondary solvent, and composition tuning, has been adjusted to the demands of the field.^{8–13} In parallel, increased device efficiencies could be achieved, e.g., by optimizing light in- and out-coupling and introducing tandem concepts.^{14,15}

Compound design requires an in-depth understanding of elementary processes occurring in organic semiconductors.¹⁶ In particular, linking the chemical structure to charge dynamics is a nontrivial task, since several factors can influence charge carrier mobility: the molecular electronic structure, the relative positions and orientations of neighboring molecules, and spatial inhomogeneities in the morphology, which determine charge carrier pathways on a macroscopic scale.¹⁷

Furthermore, the choice of the model Hamiltonian depends on the specific situation.¹⁸ For perfectly ordered defect-free crystals at low temperatures, the Drude model based on band theory¹⁹ or its extensions, which account for local electron–phonon coupling,^{20–22} are often used. At ambient conditions, however, the thermal fluctuations of the transfer integral, i.e., the nonlocal electron–phonon coupling, are on the same order of magnitude as its average value, and charge transport should be

treated as diffusion limited by thermal disorder. This can be achieved using semiclassical dynamics based on a Hamiltonian with interacting electronic and nuclear degrees of freedom.^{23–25} If nuclear dynamics are much slower than the dynamics of charge carriers and electronic coupling is weak, charge transport can be described by a Hamiltonian with static disorder, based on simple assumptions on the electronic density of states and on the hopping rates between localized states.

The latter approach is by now routinely used to study charge transport in amorphous and partially disordered small-molecule-based organic semiconductors.^{26–40} Its key ingredients are material morphology and intermolecular charge transfer (hopping) rates.⁴¹ The rates not only depend on the molecular electronic structure but are also sensitive to the mutual positions and orientations of molecules. Hence, in order to evaluate the rates, the material morphology must be known at an atomistic resolution. This can be achieved by performing molecular dynamics simulations and thus relies on force-field development for new compounds. If the required time and length scales exceed the range available to atomistic molecular dynamics, coarse-graining techniques can be used.⁴² These techniques need to be capable of back-mapping the coarse-grained representation to an atomistic resolution.

Charge transfer rates can be postulated on the basis of intuitive physical considerations, as is done in the Gaussian disorder models.^{43–46} Alternatively, charge transfer theories can be used

Received: June 8, 2011

Published: August 19, 2011

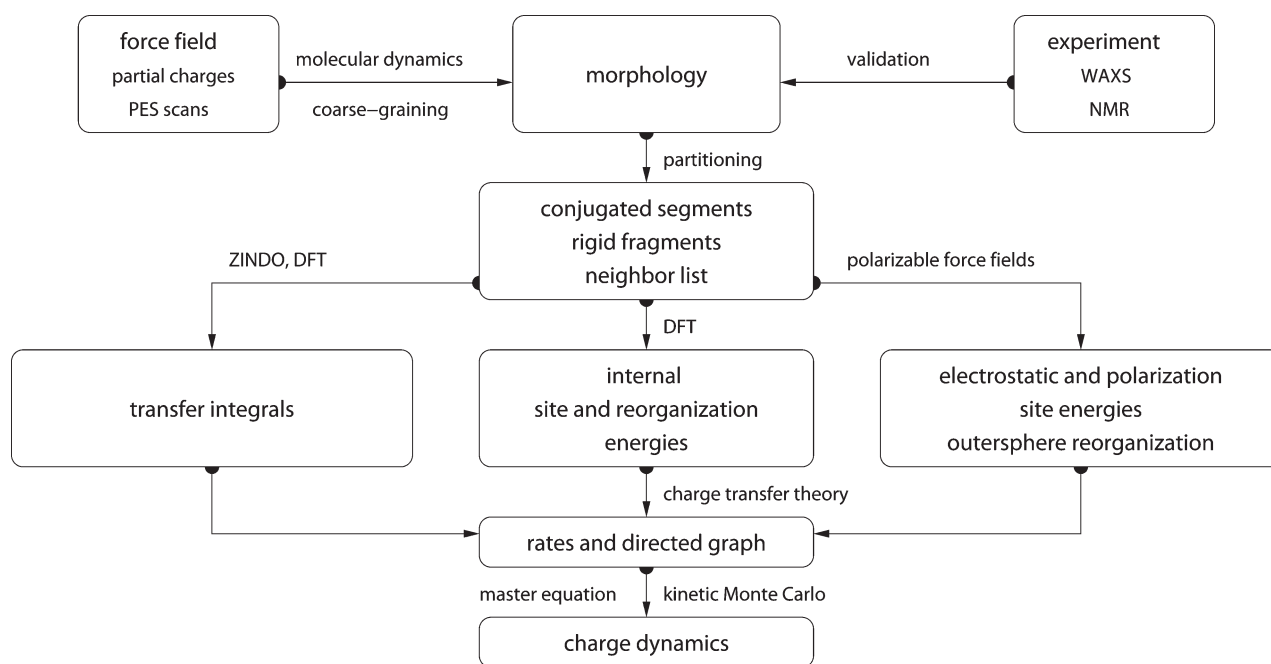


Figure 1. Workflow for microscopic simulations of charge transport. Ground state geometries, partial charges, and a refined force field are used to simulate atomistically resolved morphologies (section II.A). After partitioning into conjugated segments and rigid fragments (section II.B), a list of pairs of molecules (neighbor list) is constructed. Transfer integrals (section II.C), reorganization (section II.D) and site energies (section II.E), and eventually hopping rates are calculated for all pairs from this list. A directed graph is then generated, and the corresponding master equation is solved using the kinetic Monte Carlo method (section II.G).

to evaluate rates from quantum chemical calculations.^{28,47–51} In spite of being significantly more computationally demanding, the latter approach allows one to link the chemical and electronic structure, as well as the morphology, to charge dynamics.

The high temperature limit of classical charge transfer theory^{52,53} is often used as a trade-off between theoretical rigor and computational complexity. It captures key parameters which influence charge transport while at the same time providing an analytical expression for the rate. Within this limit, the transfer rate for a charge to hop from a site i to a site j reads:

$$\omega_{ij} = \frac{2\pi}{\hbar} \frac{J_{ij}^2}{\sqrt{4\pi\lambda_{ij}k_B T}} \exp\left[-\frac{(\Delta E_{ij} - \lambda_{ij})^2}{4\lambda_{ij}k_B T}\right] \quad (1)$$

where T is the temperature, $\lambda_{ij} = \lambda_{ij}^{\text{int}} + \lambda_{ij}^{\text{out}}$ is the reorganization energy, which is a sum of intra- and intermolecular (outer-sphere) contributions, ΔE_{ij} is the site-energy difference, or driving force, and J_{ij} is the electronic coupling element, or transfer integral.⁵⁴ A more general, quantum-classical expression for a bimolecular multichannel rate is derived in the Supporting Information.

All of the ingredients entering eq 1 can be calculated using electronic structure techniques, classical simulation methods, or their combination. With the rates at hand, one can study charge transport by solving the differential (master) equation, e.g., by using the kinetic Monte Carlo method, which is capable of simulating charge dynamics of non-steady-state systems.

Altogether, the task of charge transport characterization is rather tedious and time-consuming to perform, even for a single compound. The main aim of this work is to introduce a software package which implements a set of techniques for charge transport simulations as well as provides a flexible modular platform for their further development.

The paper is organized as follows. In section II, we describe the workflow and the basic ideas behind each method. As an illustration, we study charge transport in the bulk of amorphous tris-(8-hydroxyquinoline)aluminum (Alq_3). Section III deals with the analysis and visualization of charge dynamics. A brief summary of implementation is given in section IV.

II. METHODS

A workflow of charge transport simulations is depicted in Figure 1. The first step is the simulation of an atomistic morphology (section II.A), which is then partitioned into hopping sites (section II.B). The coordinates of the hopping sites are used to construct a list of pairs of molecules (neighbor list). For each pair, an electronic coupling element (section II.C), a reorganization energy (section II.D), a driving force (section II.E), and eventually the hopping rate are evaluated. The neighbor list and hopping rates define a directed graph. The corresponding master equation is solved using the kinetic Monte Carlo method (section II.G), which allows one to explicitly monitor the charge dynamics in the system as well as calculate time or ensemble averages of occupation probabilities, charge fluxes, correlation functions, and field-dependent mobilities (section III).

II.A. Material Morphology. There is no generic recipe on how to predict a large-scale atomistically resolved morphology of an organic semiconductor. The required methods are system-specific: for ultrapure crystals, for example, density-functional methods can be used provided the crystal structure is known from experimental results. For partially disordered organic semiconductors, however, system sizes much larger than a unit cell are required. Classical molecular dynamics or Monte Carlo techniques are then the methods of choice.

In molecular dynamics, atoms are represented by point masses which interact via empirical potentials prescribed by a force field. Force fields are parametrized for a limited set of compounds, and their refinement is often required for new molecules. In particular, special attention shall be paid to torsion potentials between successive repeat units of conjugated polymers or between functional groups and the π -conjugated system. First-principles methods can be used to characterize the missing terms of the potential energy function. The parametrization must take into account existing force-field contributions, e.g., due to nonbonded interactions. If q is the degree of freedom of interest, constrained geometry optimizations must be performed using both first principles and the force-field levels, yielding the total energies $U_{\text{fp}}(q)$ and $U_{\text{ff}}(q)$, respectively. Then, the missing force-field terms are fitted to their difference, $U_{\text{fp}}(q) - U_{\text{ff}}(q)$, using a prescribed functional form. For several identical or coupled degrees of freedom, a multidimensional fit can be used. Note that force-field validation is as important as its refinement. For instance, X-ray scattering and solid-state NMR provide information about averaged molecular arrangements to which simulation results can be compared.

As an example, the refinement of the OPLS force field for Alq₃ is described in the Supporting Information. In total, 16 bonded interactions were parametrized. To validate the force field, the glass transition temperature and bulk density of amorphous Alq₃ were compared to the experimental values. An amorphous morphology of Alq₃ was then obtained by quenching the system after equilibrating it above the glass transition temperature.

Self-assembling materials, such as soluble oligomers, discotic liquid crystals, block copolymers, partially crystalline polymers, etc., are the most complicated to study. The morphology of such systems often has several characteristic length scales and can be kinetically arrested in a thermodynamically nonequilibrium state. For such systems, the time and length scales of atomistic simulations might be insufficient to equilibrate or sample desired morphologies. In this case, systematic coarse-graining can be used to enhance sampling.⁴² Note that the coarse-grained representation must reflect the structure of the atomistic system and allow for back-mapping to the atomistic resolution.

II.B. Conjugated Segments and Rigid Fragments. With the morphology at hand, the next step is the construction of the effective electronic Hamiltonian of the system. In a static disorder approximation, this is equivalent to partitioning the system into hopping sites, or conjugated segments, and calculating charge transfer rates between them. Physically intuitive arguments can be used for the partitioning, which reflects the localization of the wave function of a charge. For most organic semiconductors, the molecular architecture includes relatively rigid, planar π -conjugated systems, which we will refer to as rigid fragments. A conjugated segment can contain one or more such rigid fragments, which are linked by bonded degrees of freedom. The dynamics of these degrees of freedom evolves on time scales much slower than the frequency of the internal promoting mode. In some cases, e.g., glasses, it can be “frozen” due to nonbonded interactions with the surrounding molecules.

To illustrate the concept of conjugated segments and rigid fragments, three representative molecular architectures are shown in Figure 2. The first one is a typical discotic liquid crystal, hexabenzocoronene. It consists of a conjugated core to which side chains are attached to aid self-assembly and solution processing. In this case, the orbitals localized on side chains do not participate in charge transport, and the π -conjugated system is both a rigid fragment and a conjugated segment.

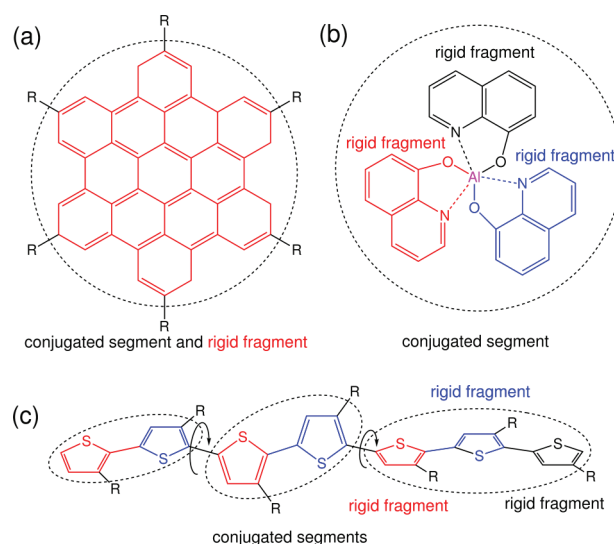


Figure 2. The concept of conjugated segments and rigid fragments. Dashed lines indicate conjugated segments, while colors denote rigid fragments. (a) Hexabenzocoronene: the π -conjugated system is both a rigid fragment and a conjugated segment. (b) Alq₃: the Al atom and each ligand are rigid fragments, while the whole molecule is a conjugated segment. (c) Polythiophene: each repeat unit is a rigid fragment. A conjugated segment consists of one or more rigid fragments. One molecule can have several conjugated segments.

In Alq₃, a metal-coordinated compound, a charge carrier is delocalized over all three ligands. Hence, the whole molecule is one conjugated segment. Individual ligands are relatively rigid, while energies on the order of $k_{\text{B}}T$ are sufficient to reorient them with respect to each other. Thus, the Al atom and the three ligands are rigid fragments.

In the case of a conjugated polymer, one molecule can consist of several conjugated segments, while each backbone repeat unit is a rigid fragment. Since the conjugation along the backbone can be broken due to large out-of-plane twists between two repeat units, an empirical criterion, based on the dihedral angle, can be used to partition the backbone on conjugated segments.³⁶ However, such intuitive partitioning is, to some extent, arbitrary and shall be validated by other methods.^{55–57}

After partitioning, an additional step is often required to remove bond length fluctuations introduced by molecular dynamics simulations, since they are already integrated out in the derivation of the rate expression. This is achieved by substituting respective molecular fragments with rigid, planar π systems optimized using first-principles methods. Centers of mass and gyration tensors are used to align rigid fragments, though a custom definition of local axes is also possible. Such a procedure also minimizes discrepancies between the force-field and first-principles-based ground state geometries of conjugated segments, which might be important for calculations of electronic couplings, reorganization energies, and intramolecular driving forces.

Finally, a list of neighboring conjugated segments is constructed. Two segments are added to this list if the distance between centers of mass of any of their rigid fragments is below a certain cutoff. This allows neighbors to be selected on a criterion of minimum distance of approach rather than center of mass distance, which is useful for molecules with anisotropic shapes.

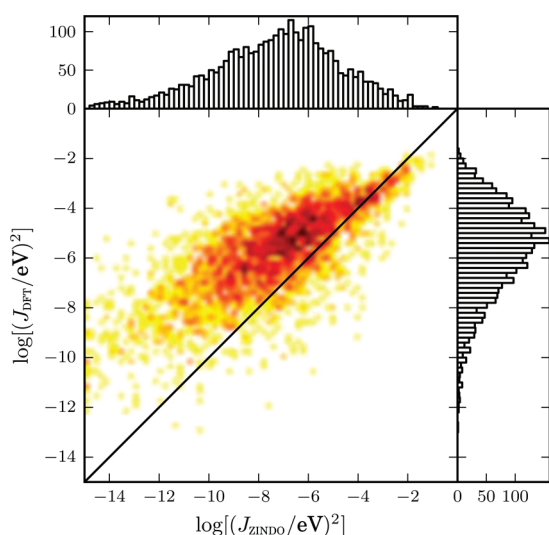


Figure 3. Distributions and correlation of transfer integrals calculated using ZINDO and DFT methods.

II.C. Transfer Integrals. The electronic transfer integral J_{ij} entering the Marcus rates in eq 1 is defined as

$$J_{ij} = \langle \phi^i | \hat{H} | \phi^j \rangle \quad (2)$$

where ϕ^i and ϕ^j are diabatic wave functions, localized on molecules i and j , respectively, participating in the charge transfer, and \hat{H} is the Hamiltonian of the formed dimer. Within the frozen-core approximation, the usual choice for the diabatic wave functions ϕ^i is the highest occupied molecular orbital (HOMO) in the case of hole transfer and the lowest unoccupied molecular orbital (LUMO) in the case of electron transfer, while \hat{H} is an effective single particle Hamiltonian, e.g., a Fock or Kohn–Sham operator of the dimer. As such, J_{ij} is a measure of the strength of the electronic coupling of the frontier orbitals of monomers mediated by the dimer interactions. Intrinsically, the transfer integral is very sensitive to the molecular arrangement, i.e., the distance and the mutual orientation of the molecules participating in charge transport. Since this arrangement can also be significantly influenced by static and/or dynamic disorder,^{24,31,34,35,53} it is essential to calculate J_{ij} explicitly for each hopping pair within a realistic morphology. Considering that the number of dimers for which eq 2 has to be evaluated is proportional to the number of molecules times their coordination number, computationally efficient and at the same time quantitatively reliable schemes are required.

In general, information about three objects is needed: the two monomer wave functions ϕ^i and ϕ^j and the dimer interaction Hamiltonian \hat{H} . An approximate method based on Zerner's independent neglect of differential overlap (ZINDO) has been described in ref 51. This semiempirical method is substantially faster than first-principles approaches, since it avoids the self-consistent calculations on each individual monomer and dimer. This allows one to construct the matrix elements of the ZINDO Hamiltonian of the dimer from the weighted overlap of molecular orbitals of the two monomers. Together with the introduction of rigid segments, only a single self-consistent calculation on one isolated conjugated segment is required. All relevant molecular overlaps can then be constructed from the obtained molecular orbitals. This molecular orbital overlap (MOO) method has been

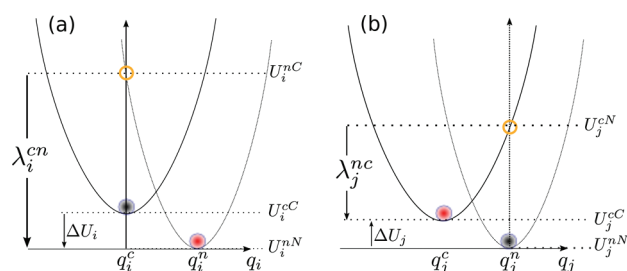


Figure 4. Potential energy surfaces of (a) donor and (b) acceptor in charged and neutral states. After the change of the charge state, both molecules relax their nuclear coordinates. If all vibrational modes are treated classically, the total internal reorganization energy and the internal energy difference of the electron transfer reaction are $\lambda_{ij}^{\text{int}} = \lambda_i^{\text{cn}} + \lambda_j^{\text{nc}}$ and $\Delta E_{ij}^{\text{int}} = \Delta U_i - \Delta U_j$, respectively.

applied successfully to study charge transport, for instance, in discotic liquid crystals,^{31,37,38} polymers,³⁶ or partially disordered organic crystals.^{33–35}

While the use of the semiempirical ZINDO method provides an efficient on-the-fly technique to determine electronic coupling elements, it is not generally applicable to all systems. For instance, its predictive capacity with regards to atomic composition and localization behavior of orbitals within more complex structures is reduced. Moreover, transition or semimetals are often not even parametrized. In this case, *ab initio* based approaches, e.g., density-functional theory, can remedy the situation.^{50,58–62} Within the dimer projection method described in detail in ref 50, explicit quantum-chemical calculations are required for every molecule and every hopping pair in the morphology. As a consequence, this procedure is significantly more computationally demanding. The code currently contains scripts which support an evaluation of transfer integrals from quantum-chemical calculations performed with the GAUSSIAN and TURBOMOLE packages.

As an example, distributions of transfer integrals calculated using ZINDO and DFT (with the gradient-corrected B–P functional^{63,64} and a TZVP basis set⁶⁵) methods are shown in Figure 3. While both distributions are similar, ZINDO integrals are, on average, smaller than DFT ones.

II.D. Reorganization Energy. The reorganization energy λ_{ij} takes into account the change in nuclear (and dielectric) degrees of freedom as the charge moves from donor i to acceptor j . It has two contributions: intramolecular, $\lambda_{ij}^{\text{int}}$, which is due to a reorganization of nuclear coordinates of the two molecules forming the charge transfer complex, and intermolecular (outersphere), $\lambda_{ij}^{\text{out}}$, which is due to the relaxation of the environment. In what follows, we discuss how these contributions can be calculated.

II.D.1. Intramolecular Reorganization Energy. If intramolecular vibrational modes of the two molecules are treated classically, the rearrangement of their nuclear coordinates after charge transfer results in the dissipation of the internal reorganization energy, $\lambda_{ij}^{\text{int}}$. It can be computed from four points on the potential energy surfaces (PES) of both molecules in neutral and charged states, as indicated in Figure 4. Adding the contributions due to discharging of molecule i and charging of molecule j yields⁴⁹

$$\lambda_{ij}^{\text{int}} = \lambda_i^{\text{cn}} + \lambda_j^{\text{nc}} = U_i^{\text{nC}} - U_i^{\text{nN}} + U_j^{\text{cN}} - U_j^{\text{cC}} \quad (3)$$

Here, U_i^{nC} is the internal energy of the neutral molecule i in the geometry of its charged state (small n denotes the state and capital C the geometry). Similarly, U_j^{cN} is the energy of the charged molecule j in the geometry of its neutral state.⁶⁶

Note that the PESs of the donor and acceptor are not identical for chemically different compounds or for conformers of the same molecule. In this case, $\lambda_i^c \neq \lambda_j^c$ and $\lambda_i^{nc} \neq \lambda_j^{nc}$. Thus, $\lambda_{ij}^{\text{int}}$ is a property of the charge transfer complex and not of a single molecule.

In Alq₃, the three ligands can easily change their mutual orientations. Molecular conformations are then “frozen” due to nonbonded interactions in an amorphous glass. The internal energies entering eq 3 were calculated after optimizing molecular geometries of all 512 molecules in charged and neutral states with the soft degrees of freedom constrained to their average values (see the Supporting Information for details). The distribution of $\lambda_{ij}^{\text{int}}$ for holes, which is shown in the Supporting Information, is sharply peaked with a maximum at 0.21 eV and variance of 0.03 eV. Computing $\lambda_{ij}^{\text{int}}$ from the PES of two unconstrained molecules leads to a similar value of 0.23 eV. Since Alq₃ has high energetic disorder arising from its large dipole moment, this small variance in reorganization energy does not affect charge carrier mobility or Poole–Frenkel behavior.

II.D.2. Outersphere Reorganization Energy. During the charge transfer reaction, also the molecules outside the charge transfer complex reorient and polarize in order to adjust for changes in electric potential, resulting in the outersphere contribution to the reorganization energy. $\lambda_{ij}^{\text{out}}$ is particularly important if charge transfer occurs in a polarizable environment. Assuming that charge transfer is much slower than electronic polarization but much faster than nuclear rearrangement of the environment, $\lambda_{ij}^{\text{out}}$ can be calculated from the electric displacement fields created by the charge transfer complex:⁶⁷

$$\lambda_{ij}^{\text{out}} = \frac{c_p}{2\epsilon_0} \int_{V^{\text{out}}} dV [D_I(\mathbf{r}) - D_F(\mathbf{r})]^2 \quad (4)$$

where ϵ_0 is the permittivity of free space, $D_{I,F}(\mathbf{r})$ are the electric displacement fields created by the charge transfer complex in the initial (charge on molecule i) and final (charge transferred to molecule j) states, V^{out} is the volume outside the complex, and $c_p = 1/\epsilon_{\text{opt}} - 1/\epsilon_s$ is the Pekar factor, which is determined by the low (ϵ_s) and high (ϵ_{opt}) frequency dielectric permittivities.

Equation 4 can be simplified by assuming spherically symmetric charge distributions on molecules i and j with total charge e . Integration over the volume V^{out} outside of the two spheres of radii R_i and R_j centered on molecules i and j leads to the classical Marcus expression for the outersphere reorganization energy:

$$\lambda_{ij}^{\text{out}} = \frac{c_p e^2}{4\pi\epsilon_0} \left(\frac{1}{2R_i} + \frac{1}{2R_j} - \frac{1}{r_{ij}} \right) \quad (5)$$

where r_{ij} is the molecular separation. While eq 5 captures the main physics, e.g., predicts smaller outersphere reorganization energies (higher rates) for molecules at smaller separations, it often cannot provide quantitative estimates, since charge distributions are rarely spherically symmetric.

Alternatively, the displacement fields can be constructed using the atomic partial charges. The difference of the displacement fields at the position of an atom b_k outside the charge transfer complex (molecule $k \neq i, j$) can be expressed as

$$D_I(\mathbf{r}_{b_k}) - D_F(\mathbf{r}_{b_k}) = \sum_{a_i} \frac{q_{a_i}^c - q_{a_i}^n}{4\pi} \frac{(\mathbf{r}_{b_k} - \mathbf{r}_{a_i})}{|\mathbf{r}_{b_k} - \mathbf{r}_{a_i}|^3} + \sum_{a_j} \frac{q_{a_j}^n - q_{a_j}^c}{4\pi} \frac{(\mathbf{r}_{b_k} - \mathbf{r}_{a_j})}{|\mathbf{r}_{b_k} - \mathbf{r}_{a_j}|^3} \quad (6)$$

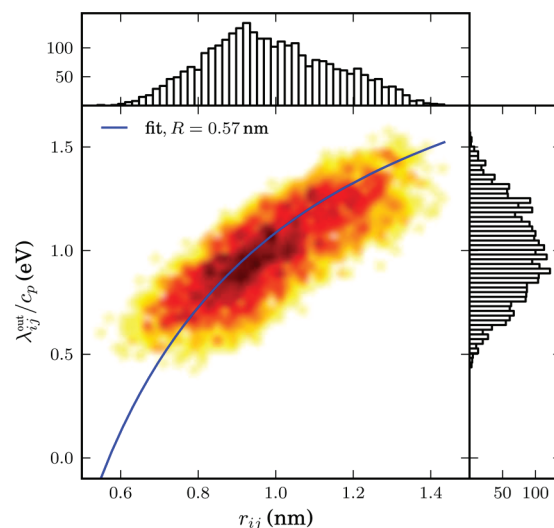


Figure 5. Outersphere reorganization energy divided by the Pekar factor as a function of the distance between two molecules and its fit to eq 5.

where q_a^n (q_a^c) is the partial charge of atom a of the neutral (charged) molecule i in a vacuum. The partial charges of neutral and charged molecules are obtained by fitting their values to reproduce the electrostatic potential of a single molecule (charged or neutral) in a vacuum. Assuming a uniform density of atoms, the integration in eq 4 can be rewritten as a density-weighted sum over all atoms excluding those of the charge transfer complex.

Using eq 6, $\lambda_{ij}^{\text{out}}/c_p$ was calculated for all pairs from the neighbor list of a system of 512 Alq₃ molecules. The neighbor list was constructed using a cutoff of 0.8 nm for the centers of mass of the three Alq₃ ligands, which results in an average of 12 neighbors in the first coordination shell. The electrostatic potential of a single molecule in a vacuum was calculated using the B3LYP functional and a 6-311G(d,p) basis set. The CHELPG method⁶⁸ was used to obtain atomic partial charges. The resulting distribution of $\lambda_{ij}^{\text{out}}/c_p$ is shown in Figure 5, together with a fit to eq 5. The fit yields $R_{\text{Alq}_3} = 0.57$ nm and predicts negative $\lambda_{ij}^{\text{out}}$ for separations smaller than this radius, which is unphysical.

The remaining unknown needed to calculate $\lambda_{ij}^{\text{out}}$ is the Pekar factor, c_p . In polar solvents $\epsilon_s \gg \epsilon_{\text{opt}} \sim 1$, and c_p is on the order of 1. In most organic semiconductors, however, molecular orientations are fixed, and therefore the low frequency dielectric permittivity is on the same order of magnitude as ϵ_{opt} . Hence, c_p is small, and its value is very sensitive to differences in the permittivities.

For Alq₃, $\epsilon_s = 3.0 \pm 0.3$ is the experimentally measured dielectric constant at low frequencies,⁶⁹ while at optical frequencies below electronic absorption, $\epsilon_{\text{opt}} = 2.9 \pm 0.1$.⁷⁰ Thus, $c_p = 0.01 \pm 0.04$, yielding outersphere reorganization energies of $\lambda_{ij}^{\text{out}} < 0.08$ eV, which are small compared to $\lambda_{ij}^{\text{int}}$. Similar results have been reported for other organic semiconductors and different methods for computing $\lambda_{ij}^{\text{out}}$.^{71–73}

II.E. Site Energy Difference. A charge transfer reaction between molecules i and j is driven by the site energy difference, $\Delta E_{ij} = E_i - E_j$. Since the transfer rate, ω_{ij} , depends exponentially on ΔE_{ij} (see eq 1), it is important to compute its distribution as accurately as possible. The total site energy difference has contributions due to an externally applied electric field, electrostatic

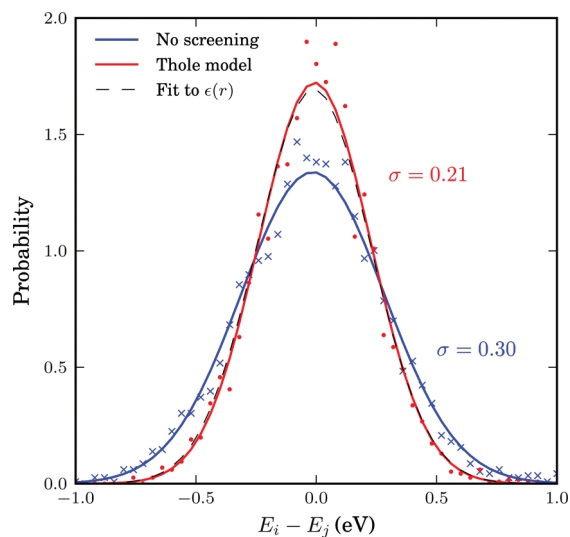


Figure 6. Distributions of site energy differences without (blue) and with (red) polarization effects for pairs from the neighbor list. Solid lines are fits to Gaussian distributions. The dashed line corresponds to a parametrized distance-dependent $\epsilon(r)$ according to eq 8.

interactions, polarization effects, and internal energy differences. In what follows, we discuss how to estimate these contributions by making use of first-principles calculations and polarizable force fields.

II.E.1. Externally Applied Electric Field. The contribution to the total site energy difference due to an external electric field F is given by $\Delta E_{ij}^{\text{ext}} = q(F \cdot r_{ij})$, where $q = \pm e$ is the charge and $r_{ij} = r_i - r_j$ is a vector connecting molecules i and j . For typical distances between small molecules, which are on the order of 1 nm, and moderate fields of $F < 10^8$ V/m, this term is always smaller than 0.1 eV.

II.E.2. Electrostatic Energy. Variations of the local electric field can result in large electrostatic contributions to the energetic disorder.⁷⁴ When the atomic partial charges of charged and neutral molecules are used, as introduced in section II.D.2, $\Delta E_{ij}^{\text{el}}$ can be computed from the site energies³¹

$$E_i^{\text{el}} = \frac{1}{4\pi\epsilon_0} \sum_{a_i} \sum_{\substack{b_k \\ k \neq i}} \frac{(q_{a_i}^c - q_{a_i}^n) q_{b_k}^n}{\epsilon_s r_{a_i b_k}} \quad (7)$$

where $r_{a_i b_k} = |r_{a_i} - r_{b_k}|$ is the distance between atoms a_i and b_k and ϵ_s is the static relative dielectric constant. The first sum extends over all atoms of molecule i , for which the site energy is calculated. The second sum reflects interactions with all atoms of neutral molecules $k \neq i$. By using eq 7, one assumes that the influence of conformational changes on partial charges and changes of the molecular geometry upon charging are small.

In order to minimize finite size effects, we do not use a spherical cutoff but apply the nearest image convention, that is, sum over all neutral molecules in the box after centering the box around the charged molecule. For Alq₃, with long-range interactions due to its large dipole moment, this procedure converges already for a few hundred molecules.

The resulting distribution of the site energy differences without screening ($\epsilon_s = 1$), shown in Figure 6, is Gaussian, with variance of $\sigma = 0.30$ eV. Note that $\Delta E_{ij}^{\text{el}}$ is constructed on the basis of the neighbor list as described in section II.D.2.

II.E.3. Polarization Effects. The influence of polarization effects on the Coulomb interactions can be taken into account by using a relative dielectric constant in eq 7. Bulk values of $\epsilon_s = 2-5$ for typical organic semiconductors uniformly scale all site energies but are not capable of describing polarization effects on a microscopic level. The contribution to E_i^{el} from the first coordination shell is then underestimated due to overscreening, and as a result, the site-energy differences become artificially small. Alternatively, one can introduce a phenomenological distance-dependent screening function $\epsilon(r_{a_i b_k})$ in eq 7³⁰

$$\epsilon(r) = \epsilon_s - (\epsilon_s - 1) \left(1 + sr + \frac{1}{2} s^2 r^2 \right) e^{-sr} \quad (8)$$

where the parameter s is the inverse screening length. For a monovalent ion in water, for example, $\epsilon_s = 80$ and $s = 3 \text{ nm}^{-1}$.⁷⁵ This screening function ensures that neighboring atoms interact via an unscreened Coulomb potential ($\epsilon \sim 1$), while the electrostatic interaction between atoms at large separations is screened as in the bulk.

While phenomenological distance-dependent screening is computationally efficient, it cannot be used for inhomogeneous systems or systems with anisotropic molecular polarizabilities. Moreover, ϵ_s and s are not known for newly synthesized compounds. A more general approach relies on self-consistent methods to obtain polarization fields.⁷⁶ Here, we use a polarizable force field based on the Thole model⁷⁷ as implemented in the TINKER package.⁷⁸

The polarization contribution is refined iteratively. After evaluating the electric field at atom a in molecule i , $F_{a_i}^{(0)}$, created by all atomic partial charges ($\epsilon_s = 1$, nearest image convention), the induced dipole moments, $\mu_{a_i}^{(0)}$, are computed. During this first step, intramolecular interactions are excluded. Induced dipole moments are then iteratively refined as $\mu_{a_i}^{(k+1)} = \omega F_{a_i}^{(k)} \alpha_{a_i} + (1 - \omega) \mu_{a_i}^{(k)}$, where α_{a_i} is the isotropic atomic polarizability and $\omega = 0.5$ is a damping constant for successive over-relaxation. The new electric fields are computed using the induced dipole moments, which now interact with each other also within molecules, allowing for anisotropic molecular polarizabilities. The procedure is repeated until $\sum_{a_i} |\mu_{a_i}^{(k+1)} - \mu_{a_i}^{(k)}| < 10^{-6}$ Debye.

Such self-consistent calculations can, however, become computationally prohibitive for large systems. For homogeneous systems and isotropic molecular polarizabilities, one can perform self-consistent calculations for small systems, parametrize eq 8 accordingly, and use $\epsilon(r)$ to study larger systems. To this end, the bulk dielectric constant is obtained from the Clausius–Mosotti relation⁷⁹

$$\epsilon_s = 1 + \frac{12\pi\alpha N/V}{3 - 4\pi\alpha N/V} \quad (9)$$

where α is the molecular polarizability volume and N/V is the number density. Using this value of ϵ_s , the parameter s is then fitted to reproduce the distribution of site-energy differences for molecules from the neighbor list.

For a neutral Alq₃ molecule, the Thole model (using atomic polarizabilities $\alpha_{\text{H,C,N,O,Al}} = 0.696, 1.75, 1.073, 0.837, 5.5 \text{ \AA}^3$, respectively, and a damping factor of $a = 0.39$ for interactions with induced moments⁷⁸) gives a practically isotropic polarizability volume tensor with $\alpha = 54.9 \text{ \AA}^3$. This agrees with DFT calculations (B3LYP functional and 6-311G(d,p) basis set), yielding $\alpha = 55.2 \text{ \AA}^3$.⁸⁰ Using $N = 512$ molecules in a cubic box of length $L = 67.8 \text{ \AA}$, we obtain $\epsilon_s = 2.84$, which reproduces the experimental value of 3.0 ± 0.3 .⁶⁹

The corresponding distributions of site energy differences, shown in Figure 6, are practically Gaussian. For the Thole model, the variance of 0.21 eV is obviously larger than the 0.10 eV obtained using bulk screening with $\epsilon_s = 3.0$ (not shown) and is smaller than 0.30 eV for $\epsilon_s = 1$. A fit to eq 8 gives $s = 1.3 \text{ nm}^{-1}$, which is significantly smaller than the inverse screening length of water, 3 nm^{-1} .

II.E.4. Internal Energy Difference. The contribution to the site energy difference due to different internal energies (see Figure 4) can be written as

$$\Delta E_{ij}^{\text{int}} = \Delta U_i - \Delta U_j = (U_i^{\text{cC}} - U_i^{\text{nN}}) - (U_j^{\text{cC}} - U_j^{\text{nN}}) \quad (10)$$

where $U_i^{\text{cC}(nN)}$ is the total energy of molecule i in the charged (neutral) state and geometry. ΔU_i corresponds to the adiabatic ionization potential (or electron affinity) of molecule i , as shown in Figure 4. For one-component systems and negligible conformational changes $\Delta E_{ij}^{\text{int}} = 0$, while it is significant for donor–acceptor systems.

In Alq₃, significant conformational changes (see section II. D.1) lead to a Gaussian distribution of $\Delta E_{ij}^{\text{int}}$ with a small variance of $\sigma^{\text{int}} = 0.01 \text{ eV}$. The internal energy disorder is small compared to the electrostatic (including polarization) energetic disorder and hence does not affect the charge carrier mobility.

II.F. Spatial Correlations of Energetic Disorder. Long-range, e.g., electrostatic and polarization, interactions often result in spatially correlated disorder,⁸¹ which affects the onset of the mobility-field (Poole–Frenkel) dependence.^{30,82,83}

To quantify the degree of correlation, one can calculate the spatial correlation function of E_i and E_j at a distance r_{ij}

$$C(r_{ij}) = \frac{\langle (E_i - \langle E \rangle)(E_j - \langle E \rangle) \rangle}{\langle (E_i - \langle E \rangle)^2 \rangle} \quad (11)$$

where $\langle E \rangle$ is the average site energy. $C(r_{ij})$ is 0 if E_i and E_j are uncorrelated and 1 if they are fully correlated. For a system of randomly oriented point dipoles, the correlation function decays as $1/r$ at large distances.⁸⁴

For systems with spatial correlations, variations in site energy differences, ΔE_{ij} , of pairs of molecules from the neighbor list are smaller than variations in site energies, E_i , of all individual molecules. Since only neighbor list pairs affect transport, the distribution of ΔE_{ij} rather than that of individual site energies, E_i , should be used to characterize energetic disorder.

For Alq₃, the spatial correlation function of the electrostatic contribution to site energies, which is calculated for 512 molecules, is shown in Figure 7. It qualitatively reveals strong correlations due to the large dipole moment of the meridional isomer of Alq₃ of approximately 4 Debye. Quantitatively, this result is not converged with respect to the system size, and bigger systems will exhibit even longer-ranged correlations. The inset of Figure 7 shows that distributions of ΔE_{ij} for all and neighbor-list-only pairs are clearly different. Note that respective distributions of *internal* site energies are identical, indicating that this type of disorder is spatially uncorrelated.

II.G. Solving the Master Equation. Having determined the list of conjugated segments (hopping sites) and charge transfer rates between them, the next task is to solve the master equation, which describes the time evolution of the system

$$\frac{\partial P_\alpha}{\partial t} = \sum_\beta P_\beta \Omega_{\beta\alpha} - \sum_\beta P_\alpha \Omega_{\alpha\beta} \quad (12)$$

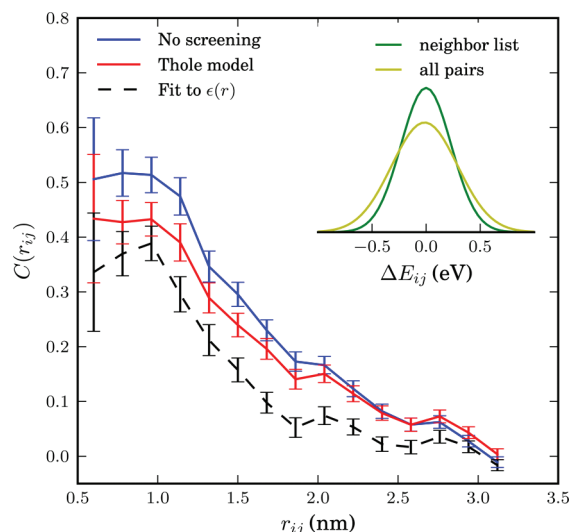


Figure 7. Electrostatic site energy correlation function (eq 11) calculated for pairs from the neighbor list without (blue) and with polarization effects from the Thole model (red) as well as using a parametrized distance-dependent $\epsilon(r)$ according to eq 8 (dashed line). Inset: Gaussian fits to electrostatic site energy distributions for all pairs ($\sigma = 0.30 \text{ eV}$) and for pairs from the neighbor list ($\sigma = 0.21 \text{ eV}$).

where P_α is the probability of the system to be in a state α at time t and $\Omega_{\alpha\beta}$ is the transition rate from state α to state β . A state α is specified by a set of site occupations, $\{\alpha_i\}$, where $\alpha_i = 1$ (0) for an occupied (unoccupied) site i , and the matrix $\hat{\Omega}$ can be constructed from rates ω_{ij} .

In particular, for a system with only one charge carrier, each state is uniquely characterized by the index i of the site the carrier occupies. In other words, only states of type $i \equiv \{0, \dots, 0, \alpha_i = 1, 0, \dots, 0\}$ are possible, and the corresponding probabilities P_i and the transition rates Ω_{ij} are identical to site occupation probabilities p_i and the transfer rates ω_{ij} , respectively. Equation 12 can then be rewritten as

$$\frac{\partial p_i}{\partial t} = \sum_j p_j \omega_{ji} - \sum_j p_i \omega_{ij} \quad (13)$$

and can be solved using linear algebra. While being efficient for stationary, low charge carrier density cases (one charge carrier per simulation box), this approach can become unstable for systems with high energetic disorder, where rates vary by several orders of magnitude.

In more general cases, such as multiple charge carriers, expressing the state picture (eq 12) in terms of site occupations is required because of an extremely large total number of states. For multiple charge carriers, the master equation can still be rewritten in terms of occupation probabilities (see the Supporting Information) by assuming only site-blocking charge–charge interactions and by using a mean-field approximation.⁸⁵ The analogue of eq 13 becomes, however, nonlinear and requires special solvers. If, in addition, several different types of carriers, such as holes, electrons, and excitons, are present in the system and their creation/annihilation processes take place, it is practically impossible to link state and site occupation probabilities and the corresponding rates.

Instead, the solution of eq 12 can be obtained by using kinetic Monte Carlo (KMC) methods. KMC explicitly simulates the

dynamics of charge carriers by constructing a Markov chain in state space and can find both stationary and transient solutions of the master equation. The main advantage of KMC is that only states with a direct link to the current state need to be considered at each step. Since these can be constructed solely from current site occupations, extensions to multiple charge carriers (without the mean-field approximation), site-occupation dependent rates (needed for the explicit treatment of Coulomb interactions), and different types of interacting particles and processes are straightforward.

To optimize memory usage and efficiency, a combination of the variable step size method⁸⁶ and the first reaction method is implemented as described in the Supporting Information.

II.H. Extrapolation to Nondispersive Mobilities. Predictions of charge-carrier mobilities in partially disordered semiconductors rely on charge transport simulations in systems which are only several nanometers thick. As a result, simulated charge transport might be dispersive for materials with large energetic disorder,^{87,88} and simulated mobilities are system-size-dependent. In time-of-flight (TOF) experiments, however, a typical sample thickness is in the micrometer range, and transport is often nondispersive. In order to link the simulation and experiment, one needs to extract the nondispersive mobility from simulations of small systems, where charge transport is dispersive at room temperature.

Such extrapolation is possible if the temperature dependence of the nondispersive mobility is known in a wide temperature range. For example, one can use analytical results derived for one-dimensional models.^{82,89,90} The mobility-temperature dependence can then be parametrized by simulating charge transport at elevated temperatures, for which transport is nondispersive even for small system sizes. This dependence can then be used to extrapolate to the nondispersive mobility at room temperature.³²

For Alq₃, the charge carrier mobility of a periodic system of 512 molecules was shown to be more than 3 orders of magnitude higher than the nondispersive mobility of an infinitely large system.³² Furthermore, it was shown that the transition between the dispersive and nondispersive transport has a logarithmic dependence on the number of hopping sites N . Hence, a brute-force increase of the system size cannot resolve the problem for compounds with large energetic disorder σ , since N increases exponentially with σ^2 .

III. MACROSCOPIC OBSERVABLES

Spatial distributions of charge and current densities can provide better insight into the microscopic mechanisms of charge transport. If O is an observable which has the value O_α in a state α , its ensemble average at time t is a sum over all states weighted by the probability P_α to be in a state α at time t

$$\langle O \rangle = \sum_{\alpha} O_{\alpha} P_{\alpha} \quad (14)$$

If O does not explicitly depend on time, the time evolution of $\langle O \rangle$ can be calculated as

$$\begin{aligned} \frac{d\langle O \rangle}{dt} &= \sum_{\alpha, \beta} [P_{\beta} \Omega_{\beta\alpha} - P_{\alpha} \Omega_{\alpha\beta}] O_{\alpha} \\ &= \sum_{\alpha, \beta} P_{\beta} \Omega_{\beta\alpha} [O_{\alpha} - O_{\beta}] \end{aligned} \quad (15)$$

If averages are obtained from KMC trajectories, $P_{\alpha} = s_{\alpha}/s$, where s_{α} is the number of Markov chains ending in the state α after time t , and s is the total number of chains.

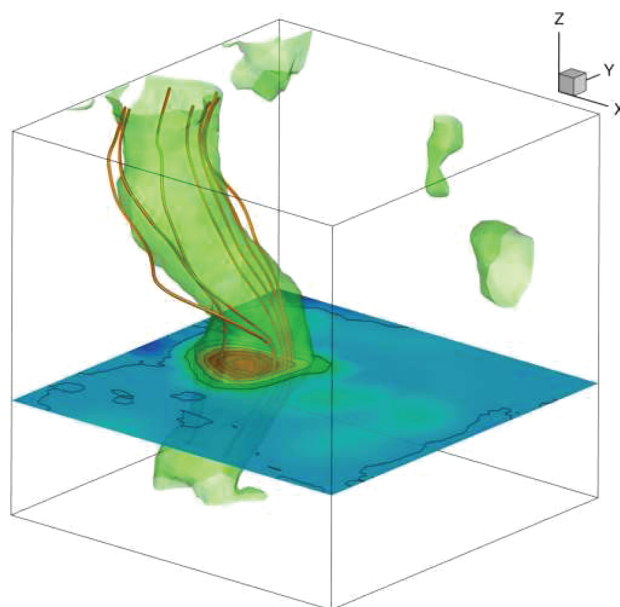


Figure 8. Isosurface of the current density for amorphous Alq₃ (512 molecules, external field $F_z = 10^7$ V/m, distance-dependent dielectric constant, DFT-based transfer integrals). Currents have filamentary structure due to large correlated energetic disorder.⁹¹ The red streamlines depict interpolated charge pathways.

Alternatively, one can calculate time averages by analyzing a single Markov chain. If the total occupation time of the state α is τ_{α} then

$$\bar{O} = \frac{1}{\tau} \sum_{\alpha} O_{\alpha} \tau_{\alpha} \quad (16)$$

where $\tau = \sum_{\alpha} \tau_{\alpha}$ is the total time used for time averaging.

For ergodic systems and sufficient sampling times, ensemble and time averages should give identical results. In many cases, the averaging procedure reflects a specific experimental technique. For example, an ensemble average over several KMC trajectories with different starting conditions corresponds to averaging over injected charge carriers in a time-of-flight experiment. In what follows, we focus on the single charge carrier (low concentration of charges) case.

III.A. Charge Density. For a specific type of particles, the microscopic charge density of a site i is proportional to the occupation probability of the site, p_i

$$\rho_i = ep_i/V_i \quad (17)$$

where, for an irregular lattice, the effective volume V_i can be obtained from a Voronoi tessellation of space. For reasonably uniform lattices (uniform site densities), this volume is almost independent of the site, and a constant volume per site, $V_i = V/N$, can be assumed. In the macroscopic limit, the charge density can be calculated using a smoothing kernel function, i.e., a distance-weighted average over multiple sites. Site occupations p_i can be obtained from eq 14 or eq 16 by using the occupation of site i in state α as an observable.

If the system is in thermodynamic equilibrium, that is without sources or sinks and without circular currents (and therefore no net flux), a condition known as detailed balance holds

$$p_j \omega_{ji} = p_i \omega_{ij} \quad (18)$$

It can be used to test whether the system is ergodic or not by correlating $\log p_i$ and the site energy E_i . Indeed, if $\lambda_{ij} = \lambda_{ji}$, the ratios of forward and backward rates are determined solely by the energetic disorder, $\omega_{ji}/\omega_{ij} = \exp(-\Delta E_{ij}/k_B T)$ (see eq 1).

III.B. Current. If the position of the charge, r , is an observable, the time evolution of its average $\langle r \rangle$ is the total current in the system

$$J = e\langle v \rangle = e \frac{d\langle r \rangle}{dt} = e \sum_{i,j} p_j \omega_{ji} (r_i - r_j) \quad (19)$$

Symmetrizing this expression, we obtain

$$J = \frac{1}{2} e \sum_{i,j} (p_j \omega_{ji} - p_i \omega_{ij}) r_{ij} \quad (20)$$

where $r_{ij} = r_i - r_j$. Symmetrization ensures equal flux splitting between neighboring sites and the absence of local average fluxes in equilibrium. It allows one to define a local current through site i as

$$J_i = \frac{1}{2} e \sum_j (p_j \omega_{ji} - p_i \omega_{ij}) r_{ij} \quad (21)$$

A large value of the local current indicates that the site contributes considerably to the total current. A collection of such sites thus represents most favorable charge pathways.

Figure 8 illustrates site currents for amorphous Alq₃ for a system of 512 molecules. The distribution of currents is very inhomogeneous, and some pathways are sampled more frequently than the others, which is a direct consequence of a rough and correlated energy landscape.⁹¹

III.C. Mobility and Diffusion Constant. For a single particle, e.g., a charge or an exciton, a zero-field mobility can be determined by studying particle diffusion in the absence of external fields. Using the particle displacement squared, Δr_i^2 , as an observable, we obtain

$$\begin{aligned} 2dD_{\gamma\delta} &= \frac{d\langle \Delta r_{i,\gamma} \Delta r_{i,\delta} \rangle}{dt} \\ &= \sum_{\substack{i,j \\ i \neq j}} p_j \omega_{ji} (\Delta r_{i,\gamma} \Delta r_{i,\delta} - \Delta r_{j,\gamma} \Delta r_{j,\delta}) \\ &= \sum_{\substack{i,j \\ i \neq j}} p_j \omega_{ji} (r_{i,\gamma} r_{i,\delta} - r_{j,\gamma} r_{j,\delta}) \end{aligned} \quad (22)$$

Here, r_i is the coordinate of the site i ; $D_{\gamma\delta}$ is the diffusion tensor, $\gamma, \delta = x, y, z$; and $d = 3$ is the system dimension. Using the Einstein relation

$$D_{\gamma\delta} = k_B T \mu_{\gamma\delta} \quad (23)$$

one can, in principle, obtain the zero-field mobility tensor $\mu_{\gamma\delta}$. Equation 22, however, does not take into account the use of periodic boundary conditions when simulating charge dynamics. In this case, the simulated occupation probabilities can be compared to the solution of the Smoluchowski equation with periodic boundary conditions (see the Supporting Information for details).

Alternatively, one can directly analyze time-evolution of the KMC trajectory and obtain the diffusion tensor from a linear fit to the mean square displacement, $\Delta r_{i,\gamma} \Delta r_{i,\delta} = 2dD_{\gamma\delta} t$.

The charge carrier mobility tensor, $\hat{\mu}$, for any value of the external field can be determined either from the average charge velocity defined in eq 19

$$\langle v \rangle = \sum_{i,j} p_j \omega_{ji} (r_i - r_j) = \hat{\mu} F \quad (24)$$

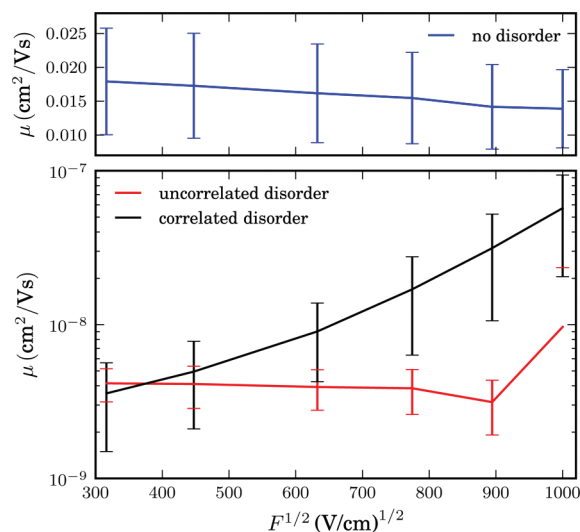


Figure 9. Poole–Frenkel plots for a system of 4096 molecules. Mobilities were calculated by averaging over ten, 0.1-s-long (10^{-5} s for no disorder), KMC runs and six different field directions. Transfer integrals were calculated using DFT; energetic disorder is based on the distance-dependent dielectric constant fitted to the site energy distribution of the Thole model.

or directly from the KMC trajectory. In the latter case, the velocity is calculated from the unwrapped (if periodic boundary conditions are used) charge displacement vector divided by the total simulation time. Projecting this velocity on the direction of the field F yields the charge carrier mobility in this particular direction. In order to improve statistics, mobilities can be averaged over several KMC trajectories and MD snapshots.

For Alq₃, the field dependence of the mobility (Poole–Frenkel plot) is shown in Figure 9 for a system of 4096 molecules. To illustrate the role of disorder and correlations, we also show the field dependence for a system without energetic disorder (top panel) and without correlated energetic disorder (randomly shuffled site energies). Energetic disorder reduces the value of mobility by 6 orders of magnitude. The Poole–Frenkel behavior for small fields can only be observed if correlated disorder is taken into account. Note that, for a system with such large energetic disorder, the absolute values of (nondispersive) mobility are systematically overestimated due to significant finite size effects (see section II.H and ref 32). The experimentally measured value of the hole mobility at small fields lies between 10^{-9} and $10^{-8} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$.⁹²

IV. IMPLEMENTATION

The toolkit is implemented using modular concepts introduced earlier in the versatile object-oriented toolkit for coarse-graining applications (VOTCA).⁴² The VOTCA structures are adapted to reading atomistic trajectories, mapping them onto conjugated segments and rigid fragments, and substituting (if needed) rigid fragments with the optimized copies.

The molecular orbital overlap module calculates electronic coupling elements between conjugated segments from the corresponding molecular orbitals. It relies on the semiempirical INDO Hamiltonian and molecular orbitals in the format provided by the GAUSSIAN package. An alternative, density-functional-based approach, has interfaces to the GAUSSIAN and TURBOMOLE packages. An interface to the TINKER package

is provided for calculations of electrostatic and polarization contributions to energetic disorder.

The kinetic Monte Carlo module reads in the neighbor list, site coordinates, and hopping rates and performs charge dynamics simulations using either periodic boundary conditions or charge sources and sinks.

The toolkit is written as a combination of modular C++ code and scripts. The data transfer between programs is implemented via a state file or database, which is also used to restart simulations. Analysis functions and most of the calculation routines are encapsulated by using the observer pattern,⁹³ which allows the implementation of new functions as individual modules.

V. SUMMARY

To summarize, we have presented a toolkit for developing and testing methods for charge transport simulations in disordered organic semiconductors. The core of the toolkit is based on a reader and a postprocessor of atomistic trajectories, a fast molecular orbital overlap calculations library, and a kinetic Monte Carlo code. To illustrate its functionality, we have studied charge transport in amorphous tris-(8-hydroxyquinoline)aluminum, a typical organic semiconductor. The source code of the toolkit is available under the Apache license (www.votca.org).

■ ASSOCIATED CONTENT

S **Supporting Information.** Derivation of the bimolecular electron transfer rate, force-field parameters, estimations of site energy differences due to conformational disorder, derivation of the master equation for multiple charge carriers, implementation of the kinetic Monte Carlo algorithm, and diffusion in periodic boundary conditions. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: denis.andrienko@mpip-mainz.mpg.de.

■ ACKNOWLEDGMENT

This work was partially supported by DFG via the IRTG program between Germany and Korea, DFG grants AN 680/1-1 and SPP1355, and BMBF grant MESOMERIE. The POLYMAT graduate program (V.R.) and Eurosime Early Stage Training project of Marie Curie actions (A.L.) are acknowledged for financial support. We acknowledge stimulating discussions with Christoph Junghans, Burkhard Dünweg, Roman Schmitz, Christian Lennartz, Andreas Fuchs, and Yuki Nagata and are indebted to J. W. Draussen-Brenner for distilling this work. We are grateful to Luigi Delle Site and Mara Jochum for a critical reading of the manuscript.

■ REFERENCES

- (1) Sommer, M.; Hüttner, S.; Thelakkat, M. *J. Mater. Chem.* **2010**, *20*, 10788.
- (2) Arias, A. C.; MacKenzie, J. D.; McCulloch, I.; Rivnay, J.; Salleo, A. *Chem. Rev.* **2010**, *110*, 3–24.
- (3) Mishra, A.; Ma, C.; Bäuerle, P. *Chem. Rev.* **2009**, *109*, 1141–1276.
- (4) Wu, J.; Pisula, W.; Müllen, K. *Chem. Rev.* **2007**, *107*, 718–747.
- (5) Allard, S.; Forster, M.; Souharce, B.; Thiem, H.; Scherf, U. *Angew. Chem., Int. Ed.* **2008**, *47*, 4070–4098.

- (6) Anthony, J. E. *Chem. Rev.* **2006**, *106*, 5028–5048.
- (7) Anthony, J. E.; Facchetti, A.; Heeney, M.; Marder, S. R.; Zhan, X. *Adv. Mater.* **2010**, *22*, 3876–3892.
- (8) Walzer, K.; Maennig, B.; Pfeiffer, M.; Leo, K. *Chem. Rev.* **2007**, *107*, 1233–1271.
- (9) Moulé, A. J.; Meerholz, K. *Adv. Funct. Mater.* **2009**, *19*, 3028–3036.
- (10) McCulloch, I.; Heeney, M.; Bailey, C.; Genevicius, K.; MacDonald, I.; Shkunov, M.; Sparrowe, D.; Tierney, S.; Wagner, R.; Zhang, W.; Chabinyc, M. L.; Kline, R. J.; McGehee, M. D.; Toney, M. F. *Nat. Mater.* **2006**, *5*, 328–333.
- (11) Müller, C.; Ferenczi, T. A. M.; Campoy-Quiles, M.; Frost, J. M.; Bradley, D. D. C.; Smith, P.; Stingelin-Stutzmann, N.; Nelson, J. *Adv. Mater.* **2008**, *20*, 3510–3515.
- (12) Zen, A.; Pflaum, J.; Hirschmann, S.; Zhuang, W.; Jaiser, F.; Asawapirom, U.; Rabe, J. P.; Scherf, U.; Neher, D. *Adv. Funct. Mater.* **2004**, *14*, 757–764.
- (13) Kline, R.; McGehee, M.; Kadnikova, E.; Liu, J.; Fréchet, J. *Adv. Mater.* **2003**, *15*, 1519–1522.
- (14) Sirringhaus, H.; Brown, P. J.; Friend, R. H.; Nielsen, M. M.; Bechgaard, K.; Langeveld-Voss, B. M. W.; Spiering, A. J. H.; Janssen, R. A. J.; Mei-jer, E. W.; Herwig, P.; de Leeuw, D. M. *Nature* **1999**, *401*, 685–688.
- (15) Rand, B. P.; Genoe, J.; Heremans, P.; Poortmans, J. *Prog. Photovoltaics* **2007**, *15*, 659–676.
- (16) Coakley, K. M.; McGehee, M. D. *Chem. Mater.* **2004**, *16*, 4533–4542.
- (17) Grozema, F. C.; Siebbeles, L. D. A. *Int. Rev. Phys. Chem.* **2008**, *27*, 87–138.
- (18) Stafström, S. *Chem. Soc. Rev.* **2010**, *39*, 2484.
- (19) da Silva Filho, D. A.; Kim, E.; Brédas, J. *Adv. Mater.* **2005**, *17*, 1072–1076.
- (20) Silinsh, E. A.; Capek, V. *Organic Molecular Crystals: Interaction Localization, and Transport Phenomena*, 1st ed.; American Institute of Physics: College Park, MD, 1997.
- (21) Silbey, R.; Munn, R. W. *J. Chem. Phys.* **1980**, *72*, 2763.
- (22) Hannewald, K.; Stojanović, V. M.; Schellekens, J. M. T.; Bobbert, P. A.; Kresse, G.; Hafner, J. *Phys. Rev. B* **2004**, *69*, 075211.
- (23) Cheung, D. L.; Troisi, A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5941.
- (24) Troisi, A.; Cheung, D. L.; Andrienko, D. *Phys. Rev. Lett.* **2009**, *102*, 116602.
- (25) McMahan, D. P.; Troisi, A. *ChemPhysChem* **2010**, *11*, 2067–2074.
- (26) Olivier, Y.; Muccioli, L.; Lemaire, V.; Geerts, Y. H.; Zannoni, C.; Cornil, J. *J. Phys. Chem. B* **2009**, *113*, 14102–14111.
- (27) Di Donato, E.; Fornari, R. P.; Di Motta, S.; Li, Y.; Wang, Z.; Negri, F. *J. Phys. Chem. B* **2010**, *114*, 5327–5334.
- (28) Nelson, J.; Kwiatkowski, J. J.; Kirkpatrick, J.; Frost, J. M. *Acc. Chem. Res.* **2009**, *42*, 1768–1778.
- (29) Kwiatkowski, J. J.; Nelson, J.; Li, H.; Brédas, J. L.; Wenzel, W.; Lennartz, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1852.
- (30) Nagata, Y.; Lennartz, C. *J. Chem. Phys.* **2008**, *129*, 034709.
- (31) Kirkpatrick, J.; Marcon, V.; Kremer, K.; Nelson, J.; Andrienko, D. *J. Chem. Phys.* **2008**, *129*, 094506.
- (32) Lukyanov, A.; Andrienko, D. *Phys. Rev. B* **2010**, *82*, 193202.
- (33) Vehoff, T.; Baumeier, B.; Andrienko, D. *J. Chem. Phys.* **2010**, *133*, 134901.
- (34) Vehoff, T.; Baumeier, B.; Troisi, A.; Andrienko, D. *J. Am. Chem. Soc.* **2010**, *132*, 11702–11708.
- (35) Vehoff, T.; Chung, Y. S.; Johnston, K.; Troisi, A.; Yoon, D. Y.; Andrienko, D. *J. Phys. Chem. C* **2010**, *114*, 10592–10597.
- (36) Rühle, V.; Kirkpatrick, J.; Andrienko, D. *J. Chem. Phys.* **2010**, *132*, 134103.
- (37) Marcon, V.; Breiby, D. W.; Pisula, W.; Dahl, J.; Kirkpatrick, J.; Patwardhan, S.; Grozema, F.; Andrienko, D. *J. Am. Chem. Soc.* **2009**, *131*, 11426–11432.
- (38) Feng, X.; Marcon, V.; Pisula, W.; Hansen, M. R.; Kirkpatrick, J.; Grozema, F.; Andrienko, D.; Kremer, K.; Müllen, K. *Nat. Mater.* **2009**, *8*, 421–426.

- (39) Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K.; Andrienko, D. *Phys. Rev. Lett.* **2007**, *98*, 227402.
- (40) Wang, L.; Nan, G.; Yang, X.; Peng, Q.; Li, Q.; Shuai, Z. *Chem. Soc. Rev.* **2010**, *39*, 423.
- (41) To model charge transport in polymers, polymer chains can be subdivided into conjugated segments, as discussed in section II.B. The latter can then be treated as separate molecules, with *intermolecular* rates given by eq 1. The *intrachain* electronic couplings are normally much higher than interchain ones, which violates assumptions made when deriving eq 1. A combination of adiabatic intrachain and nonadiabatic interchain rate expressions was, for example, used to study charge transport in amorphous polypyrrole.³⁶
- (42) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. *J. Chem. Theory Comput.* **2009**, *5*, 3211–3223.
- (43) Walker, A. B.; Kambili, A.; Martin, S. J. *J. Phys.-Condens. Mater.* **2002**, *14*, 9825–9876.
- (44) Bässler, H. *Phys. Status Solidi B* **1993**, *175*, 15–56.
- (45) Borsenberger, P. M.; Pautmeier, L.; Bässler, H. *J. Chem. Phys.* **1991**, *94*, 5447.
- (46) Pasveer, W. F.; Cottaar, J.; Tanase, C.; Coehoorn, R.; Bobbert, P. A.; Blom, P. W. M.; de Leeuw, D. M.; Michels, M. A. *J. Phys. Rev. Lett.* **2005**, *94*, 206601.
- (47) Brédas, J.; Norton, J. E.; Cornil, J.; Coropceanu, V. *Acc. Chem. Res.* **2009**, *42*, 1691–1699.
- (48) Coropceanu, V.; Cornil, J.; da Silva Filho, D. A.; Olivier, Y.; Silbey, R.; Brédas, J. *Chem. Rev.* **2007**, *107*, 926–952.
- (49) Brédas, J.; Beljonne, D.; Coropceanu, V.; Cornil, J. *Chem. Rev.* **2004**, *104*, 4971–5004.
- (50) Baumeier, B.; Kirkpatrick, J.; Andrienko, D. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11103.
- (51) Kirkpatrick, J. *Int. J. Quantum Chem.* **2008**, *108*, 51–56.
- (52) Marcus, R. A. *Rev. Mod. Phys.* **1993**, *65*, 599.
- (53) Hutchison, G. R.; Ratner, M. A.; Marks, T. J. *J. Am. Chem. Soc.* **2005**, *127*, 2339–2350.
- (54) The main assumptions in eq 1 are nonadiabaticity (small electronic coupling and charge transfer between two diabatic, noninteracting states) and harmonic promoting modes, which are treated classically. At ambient conditions, however, the intramolecular promoting mode, which roughly corresponds to C–C bond stretching, has a vibrational energy of $\hbar\omega \approx 0.2$ eV $\gg k_B T$ and should be treated quantum-mechanically. The outersphere (slow) mode has much lower vibrational energy than the intramolecular promoting mode and therefore can be treated classically. The weak interaction between molecules also implies that each molecule has its own, practically independent, set of quantum mechanical degrees of freedom. A rate expression for the aforementioned situation is derived in the Supporting Information. Numerical estimates show that if $\lambda_{ij}^{\text{int}} \approx \lambda_{ij}^{\text{out}}$ and $|\Delta E_{ij}| \ll \lambda_{ij}^{\text{out}}$, the rates are similar to those of eq 1. In general, there is no robust method to compute $\lambda_{ij}^{\text{out}}$,⁹⁴ and both reorganization energies are often assumed to be on the same order of magnitude. In this case, the second condition also holds, unless there are large differences in electron affinities or ionization potentials of neighboring molecules, e.g., in donor–acceptor blends.
- (55) Vukmirović, N.; Wang, L. *J. Chem. Phys.* **2008**, *128*, 121102.
- (56) Vukmirović, N.; Wang, L. *Nano Lett.* **2009**, *9*, 3996–4000.
- (57) McMahon, D. P.; Troisi, A. *Chem. Phys. Lett.* **2009**, *480*, 210–214.
- (58) Huang, J.; Kertesz, M. *Chem. Phys. Lett.* **2004**, *390*, 110–115.
- (59) Huang, J.; Kertesz, M. *J. Chem. Phys.* **2005**, *122*, 234707–9.
- (60) Valeev, E. F.; Coropceanu, V.; da Silva Filho, D. A.; Salman, S.; Brédas, J. *J. Am. Chem. Soc.* **2006**, *128*, 9882–9886.
- (61) Yin, S.; Yi, Y.; Li, Q.; Yu, G.; Liu, Y.; Shuai, Z. *J. Phys. Chem. A* **2006**, *110*, 7138–7143.
- (62) Yang, X.; Li, Q.; Shuai, Z. *Nanotechnology* **2007**, *18*, 424029.
- (63) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (64) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (65) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.
- (66) If the PESs of neutral and charged states are different for the same molecule, that is $\lambda_i^{\text{cn}} \neq \lambda_i^{\text{nc}}$, the rate for the bimolecular charge transfer is no longer a simple sum. If, as before, both modes are treated classically, the rate is an integral over the charge detachment and attachment spectrum of molecules *i* and *j*.⁹⁵ For most systems, however, the reorganization energies for charging or discharging of the same molecule do not deviate by more than a few percent, and the rate is given by eq 1.
- (67) May, V.; Kühn, O. *Charge and Energy Transfer Dynamics in Molecular Systems*, 2nd ed.; Wiley-VCH Verlag GmbH & Co. KGaA: New York, 2003.
- (68) Breneman, C. B.; Wiberg, K. E. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (69) Knapfer, M. *Appl. Phys. A: Mater. Sci. Process.* **2008**, *94*, 31–34.
- (70) Djuricic, A.; Kwong, C.; Lau, T.; Li, E.; Liu, Z.; Kwok, H.; Lam, L.; Chan, W. *Appl. Phys. A: Mater. Sci. Process.* **2003**, *76*, 219–223.
- (71) Martinelli, N. G.; Idé, J.; Sánchez-Carrera, R. S.; Coropceanu, V.; Brédas, J.; Ducasse, L.; Castet, F.; Cornil, J.; Beljonne, D. *J. Phys. Chem. C* **2010**, *114*, 20678–20685.
- (72) Norton, J. E.; Brédas, J. *J. Am. Chem. Soc.* **2008**, *130*, 12377–12384.
- (73) McMahon, D. P.; Troisi, A. *J. Phys. Chem. Lett.* **2010**, *1*, 941–946.
- (74) A typical example here is an amorphous phase of small molecules with a large dipole moment: For randomly oriented static dipoles of strength *d* on a regular lattice with spacing *a*, the variance of the individual site-energy distribution is given by $\bar{\sigma} \approx 2.35ed/(4\pi\epsilon_0\epsilon_s a^2)$,⁸¹ where ϵ_s is the relative dielectric constant.
- (75) Daggett, V.; Kollman, P. A.; Kuntz, I. D. *Biopolymers* **1991**, *31*, 285–304.
- (76) Verlaak, S.; Heremans, P. *Phys. Rev. B* **2007**, *75*, 115127.
- (77) Thole, B. *Chem. Phys.* **1981**, *59*, 341–350.
- (78) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (79) Feynman, R. P.; Leighton, R. B.; Sands, M. *Lectures on Physics: Commemorative Issue*; Addison Wesley: Reading, MA, 1971; Vol 2.
- (80) The Thole model also allows one to use different polarizabilities for the charged molecule. For the Alq₃ cation, DFT calculations predict $\alpha = 82.9$ Å³. The energetic disorder, however, does not change much if this value is used.
- (81) Dunlap, D.; Parris, P.; Kenkre, V. *Phys. Rev. Lett.* **1996**, *77*, 542–545.
- (82) Derrida, B. *J. Stat. Phys.* **1983**, *31*, 433–450.
- (83) Novikov, S. V.; Dunlap, D. H.; Kenkre, V. M.; Parris, P. E.; Vannikov, A. V. *Phys. Rev. Lett.* **1998**, *81*, 4472.
- (84) Novikov, S. V.; Vannikov, A. V. *J. Phys. Chem.* **1995**, *99*, 14573–14576.
- (85) Cottaar, J.; Bobbert, P. A. *Phys. Rev. B* **2006**, *74*, 115204.
- (86) Bortz, A. B.; Kalos, M. H.; Lebowitz, J. L. *J. Comput. Phys.* **1975**, *17*, 10–18.
- (87) Scher, H.; Montroll, E. *Phys. Rev. B* **1975**, *12*, 2455–2477.
- (88) Borsenberger, P. M.; Magin, E. H.; VanAuwaeraer, M. D.; Schryver, F. C. D. *Phys. Status Solidi A* **1993**, *140*, 9–47.
- (89) Cordes, H.; Baranovskii, S. D.; Kohary, K.; Thomas, P.; Yamasaki, S.; Hensel, F.; Wendorff, J. *Phys. Rev. B* **2001**, *63*, 094201.
- (90) Seki, K.; Tachiya, M. *Phys. Rev. B* **2001**, *65*, 014305.
- (91) van der Holst, J. J. M.; Uijtewaald, M. A.; Ramachandhran, B.; Coehoorn, R.; Bobbert, P. A.; de Wijs, G. A.; de Groot, R. A. *Phys. Rev. B* **2009**, *79*, 085203.
- (92) Fong, H. H.; So, S. K. *J. Appl. Phys.* **2006**, *100*, 094502–5.
- (93) Gamma, E.; Helm, R.; Johnson, R. E. *Design Patterns. Elements of Reusable Object-Oriented Software*, 1st ed.; Addison-Wesley Longman: Amsterdam, 1995.
- (94) Hoffman, B. M.; Ratner, M. A. *Inorg. Chim. Acta* **1996**, *243*, 233–238.
- (95) Kakitani, T.; Mataga, N. *J. Phys. Chem.* **1987**, *91*, 6277–6285.

Microscopic Symmetry Imposed by Rotational Symmetry Boundary Conditions in Molecular Dynamics Simulation

Amitava Roy and Carol Beth Post*

Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana, United States

S Supporting Information

ABSTRACT: A large number of viral capsids, as well as other macromolecular assemblies, have an icosahedral structure or structures with other rotational symmetries. This symmetry can be exploited during molecular dynamics (MD) to model in effect the full viral capsid using only a subset of primary atoms plus copies of image atoms generated from rotational symmetry boundary conditions (RSBC). A pure rotational symmetry operation results in both primary and image atoms at short-range, and within nonbonded interaction distance of each other, so that nonbonded interactions cannot be specified by the minimum image convention and explicit treatment of image atoms is required. As such, an unavoidable consequence of RSBC is that the enumeration of nonbonded interactions in regions surrounding certain rotational axes must include both a primary atom and its copied image atom, thereby imposing microscopic symmetry for some forces. We examined the possibility of artifacts arising from this imposed microscopic symmetry of RSBC using two simulation systems: a water shell and human rhinovirus 14 (HRV14) capsid with explicit water. The primary unit was a pentamer of the icosahedron, which has the advantage of direct comparison of icosahedrally equivalent spatial regions, for example regions near a 2-fold symmetry axis with imposed symmetry and a 2-fold axis without imposed symmetry. An analysis of structural and dynamic properties of water molecules and protein atoms found similar behavior near symmetry axes with imposed symmetry and where the minimum image convention fails compared with that in other regions in the simulation system, even though an excluded volume effect was detected for water molecules near the axes with imposed symmetry. These results validate the use of RSBC for icosahedral viral capsids or other rotationally symmetric systems.

1. INTRODUCTION

Thousands of viruses, including those that cause human disease, have protein shells that display icosahedral symmetry. This protein shell, or viral capsid, exhibits remarkable physical properties,^{1,2} which are necessary to promote the virus life cycle. These properties and their functional role in viral pathogenesis have long captured the interest of experimental and computational researchers.

All-atom MD simulations provide the most rigorous and detailed computational approach to investigate capsid physical behavior. An entire capsid solvated in water consists of ~2 million or more atoms. Simulation of such a large system is generally impractical even though a few notable exceptions have been reported.^{3,4} The capsid of human rhinovirus, one of the smallest RNA viruses, has ~500 000 atoms, and its diameter is ~300 Å. The computational time for MD simulation can be reduced by exploiting the rotational symmetry of viruses.^{5–8} Other macromolecular assemblies, such as the icosahedral pyruvate dehydrogenase complex,⁹ also display rotational symmetry, as symmetry often confers stability and results in economical usage of basic components.¹⁰ Simulation of these systems can also exploit rotational symmetry to reduce computation time. In the case of icosahedral viruses with three protomers in each of the 20 triangular tiles of an icosahedron, periodic boundary conditions allow simulation of as little as 1/60th of the entire capsid while retaining effects of the whole virus particle.^{5–8}

With periodic boundary conditions, a primary set of atoms is replicated to produce a neighboring set of atoms and so model the effect of a larger system. Symmetry related coordinates are generated by the transformation of atoms composing the primary

unit to image atoms in the neighboring units as follows:

$$\mathbf{x}(k') = \mathbf{R} \cdot \mathbf{x}(k) \quad (1)$$

where \mathbf{R} is one of the operators of the symmetry group. The coordinates of atom k , $\mathbf{x}(k)$, are transformed to the coordinates of the image, $\mathbf{x}(k')$.

Most uses of periodic boundary conditions in simulations model infinite systems and involve open crystallographic space group symmetry with cubic, truncated octahedron, etc. primary units; however, in the case of icosahedral viruses, closed point group symmetry is used to model a finite system. Thus, we examine here the case where \mathbf{R} involves pure rotational operators appropriate for viral capsids. This type of boundary condition is called rotational symmetry boundary conditions (RSBC).

Under RSBC, certain operators (eq 1) lead to short-range interactions between copies of both primary and image atoms near a rotational symmetry axis. Thus, as outlined in section 2.2, there are unavoidable occurrences of an atom in the primary unit near a rotational symmetry axis such that the distance to its own image atom (self-image) as well as distances to both primary and image atoms of neighboring atoms (replicate-image) are within the nonbonded interaction distance. The minimum image convention, typically used to evaluate nonbonded interactions in periodic boundary systems, fails in regions near such rotational symmetry axes. Nonbonded interactions must be determined by explicit treatment of image atoms as implemented with the

Received: February 5, 2011

Published: August 10, 2011

Table 1. Icosahedral Symmetry Operators^a

	1	2	3	4	5
1	<i>I</i> (1,1)	<i>F</i> ₁ (1,5)	<i>F</i> ₂ (1,4)	<i>F</i> ₃ (1,3)	<i>F</i> ₄ (1,2)
2	Z (2,1)	<i>ZF</i> ₁ (6,2)	<i>ZF</i> ₂ (11,4)	<i>ZF</i> ₃ (9,1)	<i>ZF</i> ₄ (5,1)
3	<i>X</i> (3,1)	<i>XF</i> ₁ (8,1)	<i>XF</i> ₂ (12,1)	<i>XF</i> ₃ (10,4)	<i>XF</i> ₄ (7,2)
4	<i>Y</i> (4,1)	<i>YF</i> ₁ (4,2)	<i>YF</i> ₂ (4,3)	<i>YF</i> ₃ (4,4)	<i>YF</i> ₄ (4,5)
5	<i>F</i> ₁ <i>Z</i> (2,5)	<i>F</i> ₁ <i>ZF</i> ₁ (6,1)	<i>F</i> ₁ <i>ZF</i> ₂ (11,3)	<i>F</i>₁<i>ZF</i>₃ (9,5)	<i>F</i> ₁ <i>ZF</i> ₄ (5,5)
6	<i>ZF</i> ₁ <i>Z</i> (5,2)	<i>ZF</i> ₁ <i>ZF</i> ₁ (2,2)	<i>ZF</i> ₁ <i>ZF</i> ₂ (6,3)	<i>ZF</i>₁<i>ZF</i>₃ (11,5)	<i>ZF</i> ₁ <i>ZF</i> ₄ (9,2)
7	<i>XF</i> ₁ <i>Z</i> (7,1)	<i>XF</i> ₁ <i>ZF</i> ₁ (3,5)	<i>XF</i> ₁ <i>ZF</i> ₂ (8,5)	<i>XF</i> ₁ <i>ZF</i> ₃ (12,5)	<i>XF</i> ₁ <i>ZF</i> ₄ (10,3)
8	<i>YF</i> ₁ <i>Z</i> (3,2)	<i>YF</i> ₁ <i>ZF</i> ₁ (8,2)	<i>YF</i> ₁ <i>ZF</i> ₂ (12,2)	<i>YF</i> ₁ <i>ZF</i> ₃ (10,5)	<i>YF</i> ₁ <i>ZF</i> ₄ (7,3)
9	<i>F</i> ₂ <i>Z</i> (2,4)	<i>F</i> ₂ <i>ZF</i> ₁ (6,5)	<i>F</i> ₂ <i>ZF</i> ₂ (11,2)	<i>F</i> ₂ <i>ZF</i> ₃ (9,4)	<i>F</i>₂<i>ZF</i>₄ (5,4)
10	<i>ZF</i> ₂ <i>Z</i> (12,4)	<i>ZF</i> ₂ <i>ZF</i> ₁ (10,2)	<i>ZF</i> ₂ <i>ZF</i> ₂ (7,5)	<i>ZF</i> ₂ <i>ZF</i> ₃ (3,4)	<i>ZF</i> ₂ <i>ZF</i> ₄ (8,4)
11	<i>XF</i> ₂ <i>Z</i> (11,1)	<i>XF</i> ₂ <i>ZF</i> ₁ (9,3)	<i>XF</i> ₂ <i>ZF</i> ₂ (5,3)	<i>XF</i> ₂ <i>ZF</i> ₃ (2,3)	<i>XF</i>₂<i>ZF</i>₄ (6,4)
12	<i>YF</i> ₂ <i>Z</i> (3,3)	<i>YF</i> ₂ <i>ZF</i> ₁ (8,3)	<i>YF</i> ₂ <i>ZF</i> ₂ (12,3)	<i>YF</i> ₂ <i>ZF</i> ₃ (10,1)	<i>YF</i> ₂ <i>ZF</i> ₄ (7,4)

^aRotational operators of icosahedral symmetry. *I* is identity. *F*_{1,2,3,4} are rotations around the 5-fold symmetry axis (*OP*; Figure 1) by 72°, 144°, 216°, and 288° respectively. *X*, *Y*, and *Z* are rotations by 180° around the *X*, *Y*, and *Z* axes, respectively. The first and second numbers inside the parentheses following the operator indicates the row and column numbers, respectively, of the inverse of that operator. Five operators in each row transform the protomer coordinates to five protomers for the pentameric primary unit used in this study. Neighboring images of the pentameric unit are generated by the five operators in boldface.

general image facility in the CHARMM program.^{11,12} The explicit treatment of image atoms allows for the correct calculation of forces in rotationally symmetric systems; duplicate forces are avoided with the proper counting of nonbonded interactions.^{11,12} While the enumeration of nonbonded interactions near such rotational symmetry axes is algorithmically correct, there is an unavoidable imposed symmetry in the forces from self-images and effects from the symmetry related to replicate images. Such microscopic symmetry is nonphysical but inherent to MD simulations under RSBC. Given the importance of RSBC for exploring molecular behavior of large systems with icosahedral symmetry, we executed an MD study to determine if artifacts on the dynamics exist with icosahedral RSBC due to the inherent microscopic symmetry on the force contributed by self-image and replicate-image interactions.

We conducted RSBC MD simulations on pure TIP3 water within a spherical shell using icosahedral symmetry and on the solvated capsid of human rhinovirus14 (HRV14). For both systems, the primary atoms in the simulation corresponded to five icosahedral asymmetric subunits or 1/12th of the icosahedron, and the symmetry-related image atoms were generated according to eq 1 from a group of five icosahedral operators. The use of this pentameric primary unit allows the operator **R** to be selected so that the minimum image convention fails at only certain rotational axes but not others (described below). Therefore, the properties assessed from regions near symmetry axes where the minimum image convention fails can be compared with those from regions near axes of the same symmetry but without imposed microscopic symmetry in the forces. In addition, the comparison is determined from one trajectory. This system is in contrast to the primary simulation unit being one asymmetric subunit, or the viral protomer equivalent of 1/60th of the icosahedron, for which symmetrical interactions would be imposed at all symmetry axes, and such a comparison would not be possible.

We show here from the analysis of a variety of physical properties that the imposed symmetry of forces has negligible influence on the resulting dynamics. An excluded volume for water molecules exists near axes with imposed symmetry; however this excluded volume did not perturb water structure or dynamics. Moreover, because protein atoms do not on average lie on a symmetry axis,

no excluded volume effects were apparent in the simulation of HRV14. On the basis of the similarity in the dynamics and structural properties for regions with and without symmetry imposed in the forces, we conclude that RSBC simulations do not suffer from artifacts due to microscopic symmetry imposed near certain rotational axes, and RSBC is a reliable approach for reducing a simulation system with rotational symmetry.

2. METHODS

2.1. Computer Simulations. All molecular dynamics (MD) calculations were performed using the CHARMM program^{11,12} with a constant volume and energy (NVE ensemble). A force switching function¹³ was used to smoothly truncate electrostatic and van der Waals nonbonded forces at a cutoff of 14.0 Å. The covalent bonds to hydrogen atoms were constrained by the SHAKE algorithm. The equations of motion were integrated using the Verlet leapfrog algorithm with a time step of 1 fs. Icosahedral boundary conditions implemented with the IMAGES facility within CHARMM were employed as described below. Image atoms within 16.0 Å of a primary atom were included in the nonbonded pair list. Updates of image and nonbonded lists were made heuristically. All of the simulations were performed using a parallel version of CHARMM 35b2 running on a supercomputing cluster containing 893 eight-core Dell 1950 processors. Each simulation was performed on single node containing eight cores that shared 16 GB of RAM.

2.2. Rotational Symmetry Boundary Conditions. The simulation system for both bulk water and the solvated HRV14 capsid was an icosahedral shell rather than a solid icosahedron. Water molecules were constrained within the outer and inner radii by applying spherical quadratic potentials¹⁴ referenced to 60 Å and 30 Å, respectively, for bulk water and 170 Å and 85 Å, respectively, for the solvated HRV14 capsid. The spherical quadratic potential was set up with a well depth of −0.25 kcal/mol at 1 Å from the reference distance followed by a smoothly rising repulsion. The lack of solvent in the central sphere of the icosahedron reduced the number of water molecules and avoided symmetrical interactions near the center.

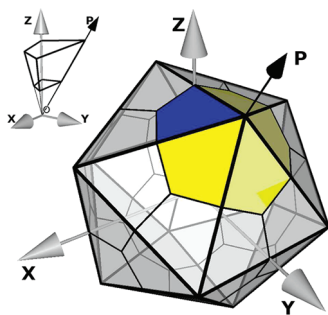


Figure 1. Schematic of an icosahedron with 20 triangular tiles. O is at the center of the icosahedron. Each tile has three asymmetric units corresponding to a protomer of the HRV14 capsid (blue area). The Z axis is coincident with one of the 2-fold symmetry axes. In this study, the pentamer, or five protomers centered around a 5-fold symmetry axis, is the primary unit cell (blue and yellow areas). The volume occupied by a protomer solvated in water, used in earlier studies,^{6–8} is shown in the inset. Here, water molecules were removed from the center of the icosahedron, and a shell of water is used to solvate a protomer (thicker lines in the inset). Symmetry operators in the first row of Table 1 generate a solvated pentamer from the solvated protomer.

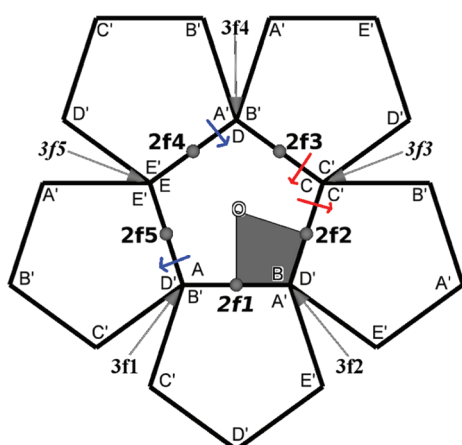


Figure 2. Two-dimensional projection of the pentamer and its neighbors generated by the set of symmetry operators shown in boldface in Table 1. The central pentamer represented by $ABCDE$, is the primary simulation unit. Pentamers represented by $A'B'C'D'E'$ are nearest-neighbor images. The area spanned by a protomer is represented by the shaded area. The 3-fold symmetry axes are $3f1$ to $3f5$, and the 2-fold symmetry axes are $2f1$ to $2f5$. O is the 5-fold symmetry axis. The Z axis is coincident with $2f1$. Self-image and replicate-image interactions are present at one 2-fold symmetry axis, $2f1$, and two 3-fold symmetry axes, $3f3$ and $3f5$.

The RSBC system was established from a subset of the 60 transformations for icosahedral symmetry listed in Table 1. The initial coordinates for the pentameric primary unit were generated by symmetry operations in the first row of Table 1 applied to a solvated HRV14 protomer (Figure 1, blue area). The solvated HRV14 protomer system, comprising four polypeptides (VP1, VP2, VP3, and VP4), was generated as described elsewhere.^{6,7} Water molecules near the center of the icosahedron were removed to model an icosahedral shell. This primary unit for the HRV14 simulation system therefore comprises five protomer units around the 5-fold symmetry axis of icosahedron (Figure 1, blue and yellow areas). Coordinates for explicit atoms in the five nearest image

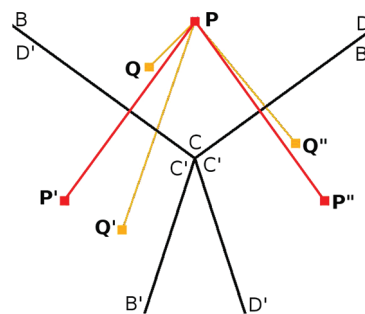


Figure 3. Illustration using the $3f3$ axis of self-image and replicate-image interactions at a 3-fold symmetry axis of an icosahedron. BCD is part of the primary unit. $B'C'D'$ are images of BCD and produced by rotating BCD around a 3-fold symmetry axis. P and Q are two atoms in primary unit and P', P'' and Q', Q'' are their images, respectively. Minimum image convention is violated as Q, Q' , and Q'' are all within nonbonded interaction distance of P (replicate-image interaction). P also interacts with its own images P' and P'' (self-image interaction). Microscopic symmetry is imposed on P as positions and velocities of P' and P'' , and Q' and Q'' are solely determined by positions and velocities of P and Q , respectively.

units were generated by transformation of the primary atomic coordinates using the five rotational operators shown in boldface in Table 1. This selection of rotational symmetry operators from Table 1 uniquely satisfies the requirements for RSBC simulation. (See the Supporting Information for a detailed explanation of the selection of symmetry operators from Table 1.) The resulting simulation system of a pentameric primary unit and the five nearest-neighbor image units is shown projected in two dimensions in Figure 2. Recall that because the central pentamer is the primary unit, each corner, labeled A through E , is distinct from the other corners. As needed for periodic boundary conditions, the crossing of a boundary between the primary unit and an image unit is equivalent for all occurrences (illustrated by the blue or red arrows). These operators generate regions with imposed microscopic symmetry in the forces from self-image and replicate-image interactions near three symmetry axes ($2f1$, $3f3$ and $3f5$), while regions around the remaining seven symmetry axes do not have such interactions. The axes with imposed symmetry are italicized through the paper to distinguish them from the other 2-fold and 3-fold axes. Self-image and replicate-image interactions are illustrated using the $3f3$ axis in Figure 3. For the pure water simulation, a pentameric primary unit and image units were similarly generated from a spatial unit of bulk water analogous to the viral protomer.

2.3. Bulk Water Simulations and Analysis. A shell of TIP3 water, with an inner and outer radius of 30 Å and 60 Å, respectively, was modeled using a primary set of atoms corresponding to a pentameric shape with 2325 water molecules, and the five nearest image sets of atoms were generated using icosahedral symmetry operators as described above. Initial velocities were assigned randomly corresponding to 100 K, and the system was heated to 300 K over a time of 20 ps and then equilibrated for 180 ps. Velocities were reassigned during the equilibration period to maintain the target temperature. Analysis was done over the subsequent 800 ps trajectory period, during which the average temperature remained constant without velocity reassignment.

The oxygen–oxygen radial distribution function for water, $g(r)$, was computed using water molecules within a certain distance of a symmetry axis for the origin of integration. The $g(r)$ was determined for water molecules surrounding five 2-fold

symmetry axes, five 3-fold symmetry axes, one 5-fold symmetry axes of the icosahedral water shell, and an arbitrary axis through the spherical water shell from an MD simulation of TIP3 water calculated without RSBC. Two cutoff distances were used to select water molecules at the origin for $g(r)$ integration: 5 and 10 Å. The average number of water molecules within 5 Å was 28 near the 3-fold axes, 43 closest to the 2-fold axes, 86 near the 5-fold axis of the icosahedral water shell, and 86 near an arbitrary axis through the spherical shell of water. Histograms of bin width 0.1 Å were used to estimate $g(r)$.

To investigate whether the symmetrical forces imposed by RSBC affect the dynamical behavior of water molecules, we calculated the survival probability function,¹⁵ $S_w(t)$, of water molecules in the vicinity of the axes of rotational symmetry. $S_w(t)$ gives the percentage of water molecules that remain in the vicinity of a symmetry axis after time t . To obtain $S_w(t)$, we compute the conditional probability $P_i(t_n, t)$ for each i th water molecule of the system. $P_i(t_n, t)$ takes a value of 1 if the i th water is in the vicinity of a symmetry axis between times t_n and $t_n + t$, and it has a value of zero otherwise. The survival probability is then

$$S_w(t) = \frac{100}{S_w(0)n_t} \sum_{n=1}^{n_t} \sum_i P_i(t_n, t) \quad (2)$$

Here, n_t is the number of simulation time-frames, and $S_w(0)$ is the average number of water molecules in the vicinity of a symmetry axis. If the oxygen atom of a water molecule is within a certain radial distance of a symmetry axis, we consider the water molecule to be within the vicinity of that symmetry axis. Image atoms were also considered while counting water molecules. With the choice of symmetry operators indicated by Figure 2, the symmetry axes 3f1, 3f2, and 3f4 correspond to equivalent regions (A, B, and D) and will therefore will have same density of water. The microenvironment and hence the number of water molecules near the 3-fold symmetry axis 3f1 will be exactly that near 3f2 and 3f4. Similarly, the count near 2-fold symmetry axis 2f2 will be exactly that near 2f3, and 2f4 will be exactly that near 2f5.

We used an orientational order parameter (Q)¹⁶ to investigate the structure of water.

$$Q \equiv 1 - \frac{3}{8} \sum_{j=1}^3 \sum_{k=j+1}^4 \left(\cos \psi_{jk} + \frac{1}{3} \right)^2 \quad (3)$$

where ψ_{jk} is the angle formed by the lines joining the oxygen atom of a given water molecule and those of nearest neighbors j and k . For the purpose of this study, we limit our attention to the four oxygen atoms nearest a given oxygen atom. Here, Q is slightly modified from previous work¹⁷ to have the value $\langle Q \rangle = 0$ ($\langle \dots \rangle$ denotes the ensemble average) in the ideal gas phase¹⁶ and $Q = 1$ for a perfect tetrahedral configuration. Thus, Q measures the degree of tetrahedrality in the distribution of the four nearest oxygen atoms around a central oxygen atom.

We considered effects on the instantaneous temperature, which depends on the root-mean-square speed, (V_{rms}).

$$V_{\text{rms}} \equiv \sqrt{\frac{\sum_{i=1}^n m_i v_i^2}{\sum_{i=1}^n m_i}} \quad (4)$$

where n is number of atoms and m_i and v_i are the mass and magnitude of velocity of the i th atom.

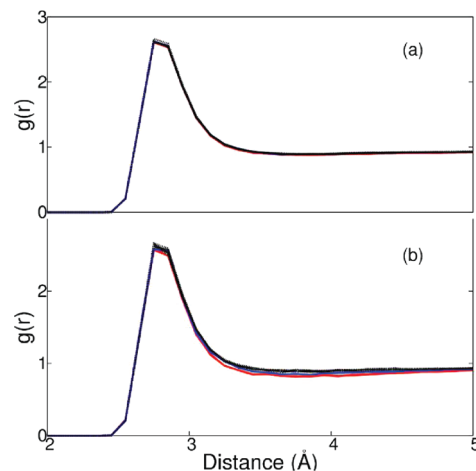


Figure 4. Radial distribution function of O–O distances for water molecules within (a) 10 Å or (b) 5 Å of 3f3 and 3f5 (red); 2f1 (blue); other 2-fold, 3-fold, and 5-fold symmetry axes of the icosahedron shell of water; and an arbitrary axis through the spherical shell of water (black).

2.4. Solvated HRV14 Capsid Simulations. Coordinates of the HRV14 protomer (protein data bank: 4RHV¹⁸) were used to generate the pentamer composing the primary set of atoms with the operators of row 1, Table 1 as described above. Missing virus coordinates were modeled on the basis of homologous residues in human rhinovirus 16 (protein data bank: 1AYM¹⁹) or polio virus (protein data bank: 1HXS²⁰). The energy of the modeled residues was minimized while maintaining the known HRV14 crystallographic coordinates at a fixed value. Keeping the known atomic positions fixed, in a vacuum, NVE MD simulations were performed to raise the temperature of the modeled regions over a 490 ps period from 100 to 5000 K; the system was annealed at 5000 K for a 50 ps period and then cooled to 300 K over a 980 ps period.

Electron density at the 3-fold and 5-fold symmetry axes has been interpreted to be a calcium ion.²¹ Following simulated annealing, one calcium ion was therefore placed at the crystallographic position on a 5-fold symmetry axis and 3-fold symmetry axis 3f1 in Figure 2. Symmetry operations generated images of a calcium ion on 3f2 and 3f4 from a calcium ion on 3f1. Modified calcium ions, with one-third of the charge and the mass of a calcium ion as described elsewhere,⁶ were placed at their crystallographic position on the 3-fold symmetry axes 3f3 and 3f5. The HRV14 pentamer with calcium ions was solvated following a known protocol. The charge of the solvated pentamer was neutralized by randomly replacing the appropriate number of water molecules with counterions, and the interior and exterior of the viral capsid were neutralized independently. The energy of the final primary unit, comprising 25 596 water molecules and a total of 142 456 atoms, was minimized using the protocol described elsewhere.⁶ Velocities corresponding to 100 K were assigned randomly to the charge neutralized viral capsid and heated to 300 K over a time of 20 ps. The solvated capsid was then allowed to equilibrate over a 980 ps period. A further 2 ns period of MD simulation was performed for analysis. Properties investigated in this article depend on the magnitude of atomic velocities, the density of atoms within a shell of different thickness near a symmetry axis, and the distribution of ϕ and ψ dihedral angles. All of these values were well converged during the 2 ns MD production period (see the Supporting Information).

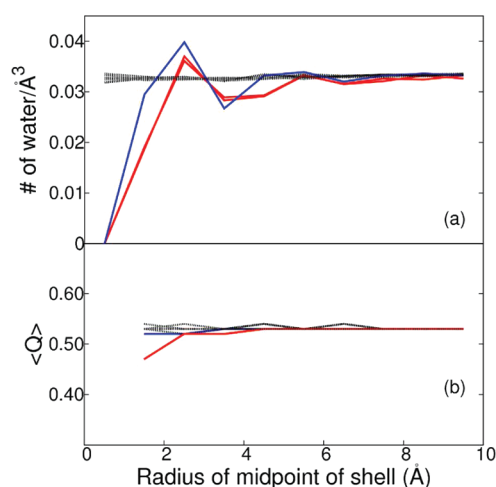


Figure 5. Density and orientational order parameter, Q , of water within cylindrical shells of 1 Å width around the $3f3$ and $3f5$ (red); $2f1$ (blue); other 2-fold, 3-fold, and 5-fold symmetry axes of the icosahedron shell of water; and an arbitrary axis through the spherical shell of water (black). (a) The density of water at X Å is calculated from the average number of water molecules between $X - 0.5$ Å and $X + 0.5$ Å over the simulation time period for X from 0.5 to 9.5 Å in 1 Å intervals. (b) $\langle Q \rangle$ within a cylindrical shell of 1 Å width is calculated by averaging the Q of individual waters within the shell over the simulation time period.

3. RESULTS AND DISCUSSION

3.1. Bulk Water. A shell of bulk water was simulated under RSBC using TIP3 water. Icosahedral symmetry restraints were set up as shown in Figure 2 with the primary unit equivalent to an icosahedral pentamer and the surrounding five replicated image units. Explicit image atoms were generated according to the operators shown in Figure 2. Symmetry is imposed on the forces from self-image and replicate-image interactions that occur near the axes $2f1$, $3f3$, and $3f5$. Interactions at the other 2-fold and 3-fold axes are uniquely determined and without imposed symmetry. Potential effects due to the nonphysical microscopic symmetry imposed $2f1$, $3f3$, and $3f5$ in RSBC MD simulations of bulk water were assessed by comparing structural and dynamic properties evaluated from water molecules near these symmetry axes with the same properties evaluated near other icosahedral symmetry axes.

3.1.1. Influence on Structural Properties. The $g(r)$ functions are plotted for averages over water molecules within 10 Å (Figure 4a) or 5 Å (Figure 4b) of a symmetry axis. The plots in Figure 4a are indistinguishable for distributions within 10 Å of any of the 12 axes, including those with imposed microscopic symmetry. Considering distributions for water molecules positioned within 5 Å of an axis, $g(r)$ is only somewhat perturbed by the imposed symmetry. Between 3 and 5 Å, the density is slightly lower for regions near $2f1$ (blue) and $3f3$ and $3f5$ (red) than regions near the other symmetry axes (Figure 4b).

An inevitable consequence of the imposed symmetry surrounding the $2f1$, $3f3$, and $3f5$ axes is that an atom cannot occupy this region in space due to steric collision with its own image. As a result, an excluded volume of cylindrical shape surrounds $2f1$, $3f3$, and $3f5$ in the simulation system. The excluded volume was characterized by determining the water density in 1-Å-thick cylindrical shells from 0.5 to 9.5 Å radii, in 1 Å increments, around each of the 2-fold, 3-fold, and 5-fold symmetry axes of the RSBC

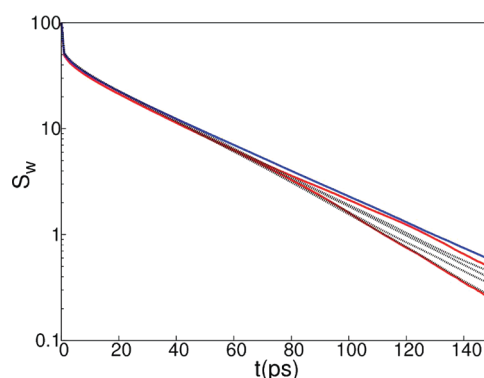


Figure 6. Survival probabilities of water molecules within 10 Å of $3f3$ and $3f5$ (red); $2f1$ (blue); and other 2-fold, 3-fold, and 5-fold symmetry axes (black) from the simulation of an icosahedron shell of water.

simulation of a water shell and around an arbitrary axis for the spherical water shell simulation. The coordinate of the water oxygen was used to specify occupancy. The water density averaged over the simulation time period is plotted as a function of the cylindrical radius in Figure 5a for each of the 12 axes defined in the Methods section. There is no water molecule within 1 Å of the $2f1$, $3f3$, or $3f5$ axis. The radius of the excluded cylindrical volume is between 1 and 2 Å and is slightly larger for the 3-fold symmetry axes than the 2-fold symmetry axis. The total excluded volume is about 0.01% of the icosahedral pentameric water shell. The excluded volume at the $2f1$, $3f3$, and $3f5$ axes is compensated by increased density between 2 and 3 Å of those symmetry axes, so that beyond 6–7 Å (less than the dimension of two water molecules), the density of water at $2f1$, $3f3$, and $3f5$ is unperturbed relative to that of bulk water and that near the other symmetry axes.

To examine whether the excluded volume and variation in water density (Figure 5a) affected the water structure in the vicinity of these symmetry axes, we examined the value of Q (see the Methods section). The ensemble averaged values $\langle Q \rangle$ for water molecules within cylindrical shells of varying radii are shown in Figure 5b. $\langle Q \rangle$ is between 0.52 and 0.53 for water molecules near all symmetry axes except those close to $3f3$ and $3f5$. The value for $\langle Q \rangle$ is 0.47 for averaging over water molecules within 2 Å of $3f3$ and $3f5$ but reaches 0.52 for averaging waters at a distance greater than 2 Å from those axes. Thus, only the first layer of water near $3f3$ and $3f5$ deviates slightly from the tetrahedrality of bulk water measured by Q .

3.1.2. Influence on Dynamical Properties. To investigate whether the symmetry imposed by RSBC affects the dynamical behavior of water molecules, we calculated the survival probability, $S_w(t)$, of water within 10 Å of symmetry axes (Figure 6).

The differences in the $S_w(t)$ of water, near the $2f1$, $3f3$, or $3f5$ axes relative to that for the other symmetry axes, are not greater than the differences among the other symmetry axes. The spread in curves in Figure 6 at 150 ps is less than one water molecule, and there is no clear trend in the spread that would indicate water molecules reside for longer or shorter times near the $2f1$, $3f3$, or $3f5$ axes compared to the other symmetry axes. Even though the density of water is lower near the $2f1$, $3f3$, and $3f5$ axes, the survival probability of water is not changed near these symmetry axes.

A comparison of the fluctuations in instantaneous temperature for the water molecules near the $2f1$, $3f3$, and $3f5$ axes with symmetry imposed in nonbonded forces to the fluctuations for the rest of the system should also indicate potential artifacts in

Table 2. Root Mean Square Speed (V_{rms}) and Flux of Water^a

	V_{rms} Å/ps	δV_{rms} Å/ps	$\Phi(S_1)$ 1/ps/Å ²	$\delta\Phi(S_1)$ 1/ps/Å ²	$\Phi(S_2)$ 1/ps/Å ²	$\delta\Phi(S_2)$ 1/ps/Å ²
2f1	0.32	0.02	0.00	0.02		
2f2,2f3,2f4,2f5	0.32	0.02	0.00	0.02		
3f3	0.32	0.03	0.00	0.02	0.00	0.02
3f5	0.32	0.02	0.00	0.02	0.00	0.02
3f1,3f2,3f4	0.32	0.02	0.00	0.02	0.00	0.02

^aRoot mean square speed (V_{rms}) and their fluctuations (δV_{rms}); flux through adjacent surfaces (Φ) and their fluctuations ($\delta\Phi$) of water molecules within 5 Å of a symmetry axes. For a 3-fold symmetry axis, the flux through two individual adjacent surfaces (S_1 , S_2) of the primary unit edge was determined. For a 2-fold symmetry axis, the flux through the primary unit surface (S_1) containing the symmetry axis was determined.

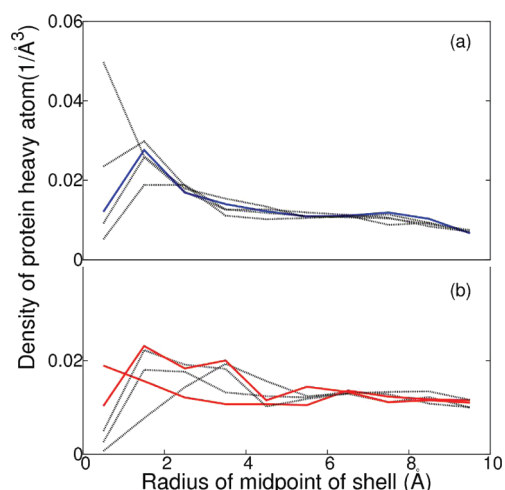


Figure 7. Density of protein within a cylindrical shell of 1 Å width around (a) 2-fold symmetry axes (2f1, blue) and other 2-fold symmetry axes (black) and (b) 3-fold symmetry axes 3f3 and 3f5 (red) and other 3-fold symmetry axes (black). The density of water at X Å is calculated by calculating the average density of protein atoms between $X - 0.5$ Å and $X + 0.5$ Å over the simulation time period. The density was calculated for X from 0.5 Å to 9.5 Å in 1 Å intervals.

the dynamical behavior of water. Fluctuations in temperature arise from fluctuations in magnitude of atomic velocities, δV_{rms} . The δV_{rms} values for water molecules in a cylinder of 5 Å radius around symmetry axes listed in Table 2 are nearly equivalent, indicating no differences in fluctuation of instantaneous temperature.

The flux, Φ , through two adjacent surfaces of the primary unit defining the corner near the 3-fold symmetry axes and the surface of the primary unit containing the 2-fold symmetry axes was also calculated. The flux does not diverge near 2f1, 3f3, and 3f5, as no net flux of water molecules was observed through the adjacent surfaces of those three symmetry axes (Table 2).

Overall, the results from MD simulation of bulk water including RSBC indicate that the symmetrical force from self-image interactions and other replicate-image nonbonded interactions imposed by an axis of rotational symmetry do not give rise to artifacts other than the presence of an excluded volume of radius between 1 and 2 Å at the symmetry axes, which alters the density of no more than two water layers. RSBC changes the structure of the first water layer somewhat near the region of excluded volume, but no differences were observed in the dynamical behavior of water molecules near these rotational axes.

3.2. HRV14 Capsid. As an asymmetric molecule, the protein atoms approach an axis of rotational symmetry but, on average, cannot lie on an axis. In the energy-minimized crystal structure,

Table 3. Root Mean Square Speed (V_{rms}) of Protein^a

3-fold symmetry axes	V_{rms} Å/ps	δV_{rms} Å/ps
3f1	0.37	0.06
3f2	0.37	0.06
3f3	0.38	0.06
3f4	0.38	0.06
3f5	0.37	0.06

^aRoot mean square speed (V_{rms}) and fluctuations (δV_{rms}) of residue 117 of VP2 near five 3-fold symmetry axes. Interactions near the symmetry axes 3f3 and 3f5 violate the minimum image convention.

the closest heavy-atom distance to a 3-fold symmetry axis is 2.5 Å, and 2.8 Å on average, while the distance to a 2-fold axis is 1.1 Å, and 1.6 Å on average. In further contrast to water molecules, given the mass and covalent structure, protein atoms do not experience overall translational motion and reorientation within the simulation boundaries. Accordingly, the protein structure and dynamics would seem less susceptible to the influence of imposed symmetry from RSBC, and effects such as the excluded volume apparent for water molecules would be diminished for a protein molecule.

The density of HRV14 atoms was determined as a function of the radial distance from the axes of rotational symmetry in the fashion outlined above for water molecules. Densities of HRV14 heavy atoms within cylindrical shells 1-Å-thick around the 2-fold and 3-fold symmetry axes are shown in Figure 7a and b, respectively. The variation among the density profiles for all of the axes is as large between those with and without symmetry imposed, and no trend in the profiles distinguishes the density of the protein atoms surrounding the 2f1, 3f3, and 3f5 axes from that surrounding the other axes. The profile around one of the 2-fold axes exhibits an interesting increase at short distance, which shows a microscopic breakdown of symmetry due to a conformational fluctuation that moves the protein toward the region of the symmetry axes. Although this type of fluctuation would be less probable at the 2f1 axis, the overall behavior shown in Figure 7 shows no differences from RSBC.

If the self-image interactions of the protein molecule near 2f1, 3f3 and 3f5 generate artificial forces, fluctuation of instantaneous temperature of the protein residues near those axes should increase. We examined the V_{rms} of residue 117 of VP2, which is nearest to the 3-fold symmetry axes. Fluctuation of V_{rms} for VP2 117 from each of the five protomers in the pentameric set of primary atoms shown in Table 3 and indicate no significant differences in trend for VP2 117 near the 3f3, 3f5 or other 3-fold symmetry axes.

We also examined structural features of residues near the five 3-fold axes. The distribution of the main chain ϕ, ψ dihedral

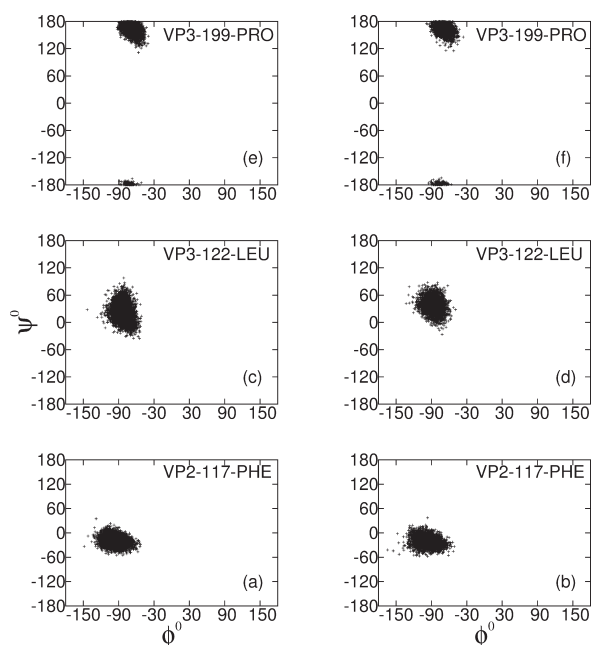


Figure 8. Ramachandran plot of residues closest to 3-fold symmetry axes. The dihedral angles, ϕ and ψ , are plotted for residue 117 of VP2 and residues 122 and 199 of VP3 from protomers located near the 3f1, 3f2, and 3f4 axes (Figure 8a, c, and e) or located near the 3f3 and 3f5 axes (Figure 8b, d, and f).

angles are shown for the residues 117 of VP2 and 122 and 199 of VP3 in Figure 8. These Ramachandran plots reveal that the distributions for these residues positioned near the 3f1, 3f2, or 3f4 axes (Figure 8a, c, and e) do not differ from those at the 3f3 or 3f5 axes (Figure 8b, d, and f), and thus the protein main-chain conformation is not effected by self-image interactions.

Together these results show that neither the structural nor dynamical behavior of the HRV viral capsid is perturbed by self-image or replicate-image interactions imposed by RSBC. The excluded volume effect, demonstrated by the density profile from bulk water simulation (Figure 5), is not manifested by the protein atoms. The fact that protein atoms, on average, cannot lie on a symmetry axis reduces the potential for artifacts induced by RSBC, and the inherent excluded volume poses practically no problem.

4. CONCLUSION

In this article, we have studied in depth the effect of the nonphysical microscopic symmetry imposed by RSBC on MD simulations of an icosahedral shell of bulk water and HRV14 viral capsid and examined the effect due to symmetry imposed at certain axes based on the similarity of properties that are well converged in the 2 ns simulation period. The approach is a reasonable critical assessment of the resulting forces and dynamics even though longer time scale properties may not be fully converged in the 2 ns simulation. Selection of the primary unit to correspond to five protomers of this virus allowed comparisons within a single trajectory of structural and dynamic properties at rotational axes with and without imposed symmetries. The only significant difference observed is the density of water within 6 Å of a rotational symmetry axis with imposed microscopic symmetry (Figure 5a). This excluded volume is unavoidable given the nonphysical nonbonded interactions to self-images. Nevertheless,

the structural and dynamical behavior of water remains without perturbation.

No differences in dynamics or structure due to the volume excluded by self-images were observed for the large protein system of HRV14. As a result of the many contributions to the atomic forces of the system, protein atoms can fluctuate into regions within 1 Å of a symmetry axis where self-image interactions occur (red line in Figure 7b). The sub-Ångstrom excluded volume due to self-image interaction affected neither the dynamics nor the structural properties of the viral protein.

In summary, we have validated the usage of RSBC to minimize the size of a system with rotational symmetry. Usage of a pentamer as the primary unit can minimize the artifacts by avoiding the higher symmetry surrounding the 5-fold axis. While a pentamer as the primary unit will only reduce the size of the system by about 12-fold instead of 60-fold if a protomer were chosen as a primary unit, a pentamer of a viral capsid, with about 65 000 atoms and 26 000 water molecules, is well within the capacity of simulation using modern day computers.

■ ASSOCIATED CONTENT

S Supporting Information. A detailed discussion of the implementation of RSBC, an explanation of minimum image violation at certain symmetry axes, and a discussion on convergence of properties investigated in the article. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cbp@purdue.edu.

Funding Sources

This work was supported by the National Institutes of Health (grant no. AI039639)

■ ACKNOWLEDGMENT

We thank Rosen Center for Advanced Computing at Purdue University for providing computing resources. We thank He Huang and Joshua Ward for meaningful discussions about symmetry operators of the reduced icosahedral group and early setup of MD simulations.

■ REFERENCES

- (1) Tama, F.; Brooks, C., III. *J. Mol. Biol.* **2005**, *345*, 299–314.
- (2) Vaidehi, N.; Goddard, W. A., III. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2466–2471.
- (3) Zink, M.; Grubmüller, H. *Biophys. J.* **2010**, *98*, 687–695.
- (4) Zink, M.; Grubmüller, H. *Biophys. J.* **2009**, *96*, 1350–1363.
- (5) Cagin, T.; Holder, M.; Pettitt, B. M. *J. Comput. Chem.* **1991**, *12*, 627–634.
- (6) Speelman, B.; Brooks, B.; Post, C. B. *Biophys. J.* **2001**, *80*, 121–129.
- (7) Li, Y.; Zhou, Z.; Post, C. B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7529–7534.
- (8) Yoneda, S.; Kitazawa, M.; Umeyama, H. *J. Comput. Chem.* **1996**, *17*, 191–203.
- (9) Milne, J. L. S.; Shi, D.; Rosenthal, P. B.; Sunshine, J. S.; Domingo, G. J.; Wu, X.; Brooks, B. R.; Perham, R. N.; Henderson, R.; Subramaniam, S. *EMBO J.* **2002**, *21*, 5587–5598.
- (10) Blundell, T. L.; Srinivasan, N. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14243–14248.

- (11) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (12) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (13) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667–683.
- (14) CHARMM:mmfp.doc. <http://www.charmm.org/documentation/c35b1/mmfp.html> (accessed July 26, 2011).
- (15) Pizzituttu, F.; Marchi, M.; Rossky, P. J.; Sterpone, F. *J. Phys. Chem. B* **2007**, *111*, 7584–7590.
- (16) Chau, P. L.; Hardwick, A. J. *Mol. Phys.* **1998**, *93*, 511–518.
- (17) Giovambattista, N.; Debenedetti, P. G.; Sciortino, F.; Stanley, H. E. *Phys. Rev. E* **2005**, *71*, 061505–061512.
- (18) Rossmann, M. G.; Arnold, E.; Erickson, J. W.; Frankenberger, E. A.; Griffith, J. P.; Hecht, H. J.; Johnson, J. E.; Kamer, G.; Luo, M.; Mosser, A. G.; Rueckert, R. R.; Sherry, B.; Vriend, G. *Nature* **1985**, *317*, 145–153.
- (19) Hadfield, A. T.; Lee, W.; Zhao, R.; Oliveira, M. A.; Minor, L.; Rueckert, R. R.; Rossmann, M. G. *Structure* **1997**, *5*, 427–441.
- (20) Miller, S. T.; Hogle, J. M.; Filman, D. J. *J. Mol. Biol.* **2001**, *307*, 499–512.
- (21) Zhao, R.; Hadfield, A. T.; Kremer, M. J.; Rossmann, M. G. *Virology* **1997**, *227*, 13–23.

Comparison of Tetrahedral Order, Liquid State Anomalies, and Hydration Behavior of mTIP3P and TIP4P Water Models

Divya Nayar, Manish Agarwal, and Charusita Chakravarty*

Department of Chemistry, Indian Institute of Technology—Delhi, New Delhi 110016, India

ABSTRACT: The relationship between local tetrahedral order, tagged particle potential energy, and coordination number is studied for mTIP3P and TIP4P models of water in the bulk as well as in the neighborhood of a small peptide. The tendency of water molecules with different binding or tagged particle potential energies to occupy environments with different degrees of disorder can be effectively illustrated by constructing tetrahedral order distributions and corresponding entropy metrics conditional on restricted ranges of local binding energy. At the state point corresponding to the onset of the density anomaly, the correlation between tetrahedral entropy versus tagged potential energy is strong and virtually identical for mTIP3P and TIP4P. In TIP4P, this correlation is retained up to temperatures as high as 300 K, while it is lost by 250 K in mTIP3P. In the 250–300 K regime that is important for biomolecular simulations, mTIP3P behaves essentially as a simple liquid while TIP4P shows the density and related anomalies characteristic of water. We also study the number of water molecules, the tetrahedral order, and the tagged molecule potential energies for water molecules as a function of the distance from the peptide for the 16-residue β -hairpin fragment of 2GB1 in mTIP3P and TIP4P solvents. The hydration shell coordination profiles ($n(r)$) of the number of water molecules are almost identical in the two solvents, but the radial variation in the local energies and local order show significant differences. The residue-wise variation in the tagged potential energy of water molecules within the first hydration shell is qualitatively similar in the two models. A comparison of the tetrahedral order distributions of water molecules lying at different distances from the biomolecular solute shows that the perturbation in the local tetrahedral order distributions of the bulk solvent due to the presence of the solute is marginal. Thus, in the 250–300 K regime, the mTIP3P and TIP4P water models show qualitatively different behavior in terms of the relationship between tetrahedral order and local energy, but as solvents in the neighborhood of a biomolecular solute, the differences between the two models are only quantitative and not qualitative.

1. INTRODUCTION

Understanding the relationship between the bulk properties of water and its behavior as a solvent is of central importance in many biomolecular and chemical processes.^{1–8} Water has a number of unusual properties that determine its behavior as a solvent. The high dielectric constant of water, associated with strong electrostatic and induction interactions, is responsible for its ability to dissolve polar and charged solutes effectively. The quantum effects associated with the very light hydrogen atoms have a significant effect on the phase diagram, equation of state, and transport properties of water. The fluctuating, three-dimensional, locally tetrahedral liquid state network, held together by hydrogen bonds, gives rise to the thermodynamic and kinetic anomalies of water and plays a crucial role in hydrophobic hydration. An understanding of the complexity of interactions in water, and their consequences for solvation, is necessary in order to understand the relationship between structure, energetics, and dynamics of solvent and solute molecules underlying hydration processes.

Liquids for which the randomly packed, hard-sphere fluid forms a reasonable zeroth-order model are classed as simple liquids.⁹ The local coordination in simple liquids is approximately icosahedral with a coordination number between 10 and 12. In contrast, tetrahedral, network-forming liquids have a local coordination of about 4 due to strongly anisotropic, local bonds.¹⁰ Tetrahedrality in water is imposed by the strong, linear hydrogen bonds between oxygen atoms of two neighboring water molecules, with each water molecule forming two donor

and two acceptor hydrogen bonds. Increased compression or heating will tend to destroy the tetrahedral bonding and shift the behavior of the system toward that of a simple liquid. Thus a tetrahedral liquid will be dominated by molecules in local environments with tetrahedral symmetry, with a fraction of nontetrahedral local environments that will depend on density and temperature. Tetrahedral liquids can therefore show a density-driven shift in the nature of local orientational order leading to water-like thermodynamic and kinetic anomalies.^{11–21} In the case of water, well-known anomalies include the rise in density on isobaric heating (density anomaly) and the increase in molecular mobility on isothermal compression (diffusivity anomaly).^{4,7} With decreasing temperature, the energetic bias toward tetrahedral order will tend to increase the fraction of four-coordinate, tetrahedral local environments. In the case of water and silicon, this temperature-dependent change in local tetrahedral order has been predicted to lead to a first-order phase transition between low- and high-density liquids at sufficiently low temperatures. The line of first-order liquid–liquid phase transitions in the pressure–temperature (PT) plane is expected to end in a second critical point.^{22–24} In the case of water, low- and high-density amorphous ice forms have been observed, although direct verification of this liquid–liquid critical point (LLCP) is difficult since the associated pressure–temperature regime is difficult to access experimentally. The predicted

Received: April 20, 2011

Published: August 15, 2011

temperature of the LLCP lies between 190 and 232 K, while the predicted pressures range from negative values to as high as 1350 atm.^{7,25} An indirect test of the presence of such a second critical point is to look for the Widom line in the experimentally accessible regime, defined as the line of maximum correlation length, or more specifically compressibility maxima, in the PT plane that terminates at the LLCP.^{22,23} It should be pointed out that while the thermodynamic and kinetic anomalies of water are well-established experimentally in the supercooled regime, experimental support for the liquid–liquid phase transition and the second critical point remains elusive.

The unusual thermodynamic properties of tetrahedral liquids, as exemplified by water, are expected to affect solvation. Hydration of small hydrophobic solutes, especially the existence of a solubility minimum as a function of temperature along an isobar, has been shown to be strongly sensitive to the equation of state and, therefore, also to the location of the density anomaly in the pressure–temperature (PT) or density–temperature plane.^{1,26,27} Given the length-scale dependence of hydrophobicity, with a shift from entropy-driven to enthalpy-driven hydration as solute size increases, the relationship between liquid-state anomalies and hydration behavior may be somewhat different for large hydrophobic solutes.^{28–31} Biomolecular solutes, such as proteins and nucleic acids, present an even more complex hydration problem because of their nanoscale size and the presence of polar and nonpolar groups. Interestingly, despite their structural complexity, both categories of biomolecules display an internal dynamical transition for temperatures lying between 180 and 220 K, associated with a sharp rise in the temperature dependence of the mean square atomic displacements (MSDs) of heavy atoms of the biomolecule.^{32–36} This dynamical transition has been shown to be largely solvent-driven and conjectured to arise from crossing of the Widom line of compressibility maxima of bulk water.³⁵ An understanding of the relationship between liquid-state anomalies and hydration behavior would therefore appear to be crucial for designing good interaction potentials for biomolecular simulations. Developing accurate as well as computationally efficient solvent models for water is, however, still a challenge because of the complex interplay of different intermolecular interactions in water like dipole–dipole, dipole–quadrupole, hydrogen-bonding, dispersion, inductive, and repulsive interactions.³⁷ While *ab initio* approaches for generating water potentials are accurate, they have so far proved cumbersome to implement in bulk simulations.^{38,39} The most widely used water models are therefore empirical rigid-body water models which freeze the intramolecular vibrations of the water molecule and model the intermolecular interactions using a distribution of Lennard-Jones and charge-carrying sites on the molecules. These water models were originally developed to reproduce the experimental properties of water under standard conditions. Recent work shows that they differ significantly in their ability to capture features of liquid state anomalies and phase diagrams.^{17,18,40–42} For example, the SPC/E model is known to reproduce the anomalous properties of water but at temperatures which are 30–40 K lower than the experimental value of 279 K at 1 atm of pressure.^{4,11,43} The version of the TIP3P potential, referred to here as modified TIP3P or mTIP3P, used with the CHARMM force field shows a TMD about 80 K lower than that found experimentally.^{41,42,44,45} Four-centered rigid-body water models, especially TIP4P and TIP4P(2005), have been shown to provide the best description of liquid water and ice.^{40,46–48} Recent studies by Nutt et al. that compare the effect of three different water models (mTIP3P, TIP4P, and TIPSP) on protein solvation dynamics and

structure are of particular interest in this context.^{49,50} They found qualitatively similar hydration behavior for all three water models, with relatively small quantitative differences. Interestingly, increasing temperature amplifies the differences between the water models for quantities such as the mean square displacements (MSDs) of heavy atom positions of the solvated protein while attenuating the differences between other quantities, such as reorientational relaxation times.⁴⁹

Since the critical factor controlling the anomalous behavior of water models is the energetic bias toward local tetrahedral order, it is important to understand the relationships between local order and local energy in the presence and absence of solutes. A previous study by Agarwal et al. studied the relationship between local tetrahedral order, local energy, and mobility in bulk water and in the hydration shell of two small peptides, using the CHARMM force field in conjunction with the TIP3P model of water.⁵¹ The local tetrahedral order metric, q_{tet} , associated with an oxygen atom i was defined as

$$q_{\text{tet}} = 1 - \frac{3}{8} \sum_{j=1}^3 \sum_{k=j+1}^4 \left(\cos \psi_{jk} + \frac{1}{3} \right)^2 \quad (1)$$

where ψ_{jk} is the angle between the bond vectors r_{ij} and r_{ik} where j and k label the four nearest oxygen atoms, so that perfect tetrahedrality corresponds to $q_{\text{tet}} = 1$.^{11,52} When determining the q_{tet} of water in the neighborhood of a solute, the definition of q_{tet} was extended to include heavy atoms of the peptide like C, O, and N as the nearest neighbors of a water molecule. As a measure of local energy, the tagged molecule potential energy (TPE or u_{tag}), defined as the interaction energy of an individual water molecule with all other molecules in the system, was used.^{53–57} The TPE can also be considered as the binding energy of an individual water molecule in the system. In the case of pair-additive potentials, the TPE can be measured as $u_{\text{tag}} = U_{\text{tot}} - U_{N-1}$ where U_{tot} is total configurational energy of the system consisting of N atoms and U_{N-1} is the configurational energy when a specific molecule is artificially removed from the configuration, keeping the positions of the rest of the atoms unchanged.⁵⁸ The Appendix contains the equations for evaluating TPE for a pair-additive intermolecular force field with Lennard-Jones and Coulombic interactions. It should be noted that, for polarizable systems, the definition of the TPE must state whether the environmental polarization is allowed to relax on removal of a given molecule.

The ensemble averages of the measures of local order and local energy were shown to be anticorrelated with each other along isochores of bulk SPC/E and mTIP3P water models.⁵¹ Water in the hydration layer of a peptide showed an oscillatory variation in the TPE with the distance from the peptide for water molecules lying between a 3 and 10 Å radius. These variations were on the order of 2–5% of the bulk TPE value and were anticorrelated with variations in local tetrahedral order in terms of locations of maxima and minima. Within a radius of 3 Å, the perturbation of the solvent structure is very significant, with local TPEs that are 10–15% lower than the bulk value and a fairly strong residue dependence.⁵¹ A recent study of the hydration shell of lysozyme confirms the anticorrelation between the TPE and q_{tet} and shows that both quantities show some sensitivity to different secondary structural units.⁵⁹ In this study, we examine the behavior of two common rigid-body water models—mTIP3P and TIP4P—with regard to the relationship between local order and energetics both in the bulk as well as in the hydration shell of a small peptide.

Table 1. Lennard-Jones Parameters (Taken from refs 44 and 45), Dipole Moment (μ), and Quadrupole Moment (Q_T) (Taken from ref 40) for Rigid-Body Water Models TIP3P, mTIP3P, and TIP4P^a

water models	ϵ_{OO}	σ_{OO}	ϵ_{HH}	σ_{HH}	q_{O}	q_{H}	q_{M}	μ	Q_T	μ/Q_T
TIP3P	0.152	3.151			-0.834	0.417		2.35	1.721	1.363
mTIP3P	0.152	3.151	0.046	0.400	-0.834	0.417		2.35	1.721	1.363
TIP4P	0.155	3.150				0.520	-1.04	2.18	2.147	1.014

^aThe units for energy, length, charge, dipole moment, quadrupole moment, and ratio μ/Q_T are kcal mol⁻¹, Å, e, Debye, Debye·Å, and Å⁻¹, respectively.

Table 2. Average Pressure Values for NVT Bulk Water Simulations of mTIP3P and TIP4P Using NAMD

water model	ρ (g cm ⁻³)	T (K)	$\langle P \rangle$ (katm)	error _P (katm)
mTIP3P	1.00	260	-0.718	0.0036
	0.95	202	-1.674	0.0031
TIP4P	1.00	260	0.074	0.0036
	0.98	257	-0.007	0.0011

These two water models have essentially the same rigid-body geometry and Lennard-Jones parameters for the oxygen–oxygen repulsion–dispersion interaction (see Table 1). The additional Lennard-Jones sites on the hydrogen atoms are relatively inconsequential for determining bulk structure and dynamics.⁶⁰ The critical difference between the two models is the location of the charged sites, which implies very different electrostatic interactions. In the case of mTIP3P and TIP4P, for example, this results in very different quadrupole moments even though dipole moments are similar.^{61,62} Since the two water models have very different temperature regimes for the anomalies, the differences in hydration behavior in the two solvent models should allow one to assess the impact of liquid-state anomalies on hydration behavior. We perform a detailed analysis of tetrahedral order distributions, local energy distributions, and coordination number for mTIP3P and TIP4P to establish the nature of correlation between these three parameters which characterize the local environment of water molecules. Though it is intuitively obvious that the tetrahedral order, local energy, and coordination number of water are related, the quantitative correlations have not been examined in great detail, and previous studies have largely focused on the ensemble-averaged values of the two quantities.⁵¹ We also perform simulations of the 16-residue β -hairpin fragment of the 2GB1 protein in both mTIP3P and TIP4P solvents. We examine the order and energetics of mTIP3P and TIP4P solvent molecules as a function of the distance from the peptide. The energetics within the hydration shell are examined by computing the TPEs in the neighborhood of different amino acid residues. We relate the differences in hydration shell structure and energetics between the two solvent models to the differences in the parameterization of the water models. Computational details of the simulations are given in section 2. Results are contained in section 3, and our conclusions are presented in section 4.

2. COMPUTATIONAL DETAILS

2.1. Bulk Water. The bulk water simulations consisted of a system of 256 molecules of TIP3P, mTIP3P, or TIP4P in cubic

simulation cells. These rigid-body potential energy surfaces for water are described by contributions from short-range interactions defined by Lennard-Jones potential and long-range electrostatic interactions which were calculated using Ewald sums. The parametric form of the interaction between two water molecules a and b is given by:

$$U_{ab} = \sum_i \sum_j \frac{q_i q_j}{r_{ij}} + 4\epsilon \left(\frac{\sigma^{12}}{r_{\text{OO}}^{12}} - \frac{\sigma^6}{r_{\text{OO}}^6} \right) \quad (2)$$

where i and j index partial charges located on molecules a and b respectively and r_{OO} refers to the distance between the oxygen atoms of the two monomers. The potential energy parameters used for TIP3P, mTIP3P, and TIP4P are given in Table 1.

For studying bulk water, we used the DL_POLY molecular dynamics package^{63,35} as well as the NAMD biomolecular suite.^{66,67} The DL_POLY package was used to ensure consistency with our previous work on water, in particular, the evaluation of the tagged molecule potential energies discussed in the Appendix.^{42,51} The NAMD package is more convenient for biomolecular simulations, and as a check on the TPE evaluation procedure using the *pairInteraction* command, we ran bulk water simulations with NAMD 2.7. All bulk water simulations were carried out in a canonical (NVT) ensemble with a time step of 1 fs, and production runs were carried out for 4 ns. The equations of motion were integrated using the Verlet leapfrog algorithm in DL_POLY simulations and velocity Verlet algorithm in NAMD simulations. The rigid-body constraints were implemented using the SHAKE algorithm in DL_POLY simulations and RATTLE algorithm in NAMD simulations. The calculation of long-range electrostatic contributions to configurational energy is done by using different methods in DL_POLY and NAMD. DL_POLY uses Ewald sums, whereas NAMD uses the particle mesh Ewald (PME) method (see Appendix A for details). The results of TPE for bulk system were found to be identical for both the cases.

Simulations of the TIP3P system were carried out for temperatures ranging from 170 K to 300 K and densities ranging from 0.90 g cm⁻³ to 1.20 g cm⁻³. Simulations for TIP4P and mTIP3P water were carried out at 1.00 g cm⁻³ density and at temperatures ranging from 200 K to 350 K. The state point of (260 K, 1.00 g cm⁻³) has been examined in detail for mTIP3P and TIP4P. For all of the state points, the compressibility was positive and the system was homogeneous, even at negative pressures, indicative of thermodynamic stability. Table 2 tabulates NVT ensemble results respectively for selected state points of bulk water. The oxygen–oxygen radial distribution functions of mTIP3P and TIP4P at 300 K and 1.00 g cm⁻³ were shown to reproduce those available in the literature.⁶⁸ Error bars on the structural and TPE distributions were found to be less than the symbol size in the associated figures.

2.2. Solvated Peptide. The solvated system of peptide was prepared by extracting the 16-residue C-terminal fragment of immunoglobulin binding domain of streptococcal protein G (PDB ID: 2GB1) and then solvating this β -hairpin fragment with 1774 TIP4P water molecules.^{51,69} The C-terminal and N-terminal residues did not carry any charges. The two aspartic acids and two glutamic acids had negative charges on their side chains, and one lysine residue carried positive charge. Thus, the solvated peptide had a net -3 charge. In order to maintain charge neutrality, three waters were replaced by three sodium ions in the system. The solvation and ionization were done in a cubic box using the “*solvate*” and “*autoionize*” plugins of VMD 1.8.7, respectively.^{70,71}

Table 3. Length of Cubic Simulation Cells and Density (ρ) of Water in System of Peptide Solvated with mTIP3P and TIP4P at 250 K and 300 K

temperature (K)	boxlength (Å)		ρ (g cm ⁻³)	
	TIP4P	mTIP3P	TIP4P	mTIP3P
250	37.89	37.09	0.973	1.020
300	38.32	37.67	0.950	0.972

Molecular dynamics simulations of the solvated peptide system were done with both mTIP3P and TIP4P using the NAMD 2.7 package. The CHARMM22 force field was used for the simulations. To compute the short-range interactions, a cutoff equal to half the box length was used. The contribution of long-range electrostatic interaction to the configurational energy was computed using the particle mesh Ewald (PME) method. Energy minimization was done using the conjugate gradient method to remove any steric repulsions among molecules of the system which may arise during the system preparation. The energy minimized system was then subjected to heating in an isobaric–isothermal (NPT) ensemble, at the rate of 1 K per 2500 steps until 250 K was reached and then 1 K per 7500 steps from 250 K to 300 K at 1 atm of pressure. The system was then equilibrated in the NPT ensemble in order to obtain the equilibrated volume value. The resulting box lengths of the cubic box of solvated peptide at 300 K and 250 K are given in Table 3. Using this equilibrated value of volume, the production runs were carried out in the microcanonical (NVE) ensemble for 2 ns using a time step of 2 fs. The configurations were stored every 100 steps (0.2 ps). The keyword *pairInteraction* was used to calculate the tagged potential energy of a given water molecule with all other molecules in the solvated peptide system. Note that the time step is consistent with that used in previous studies.⁶⁹ Tables 4 and 5 show the NVE simulation results for pressure and temperature with associated checks on the error; the drift in temperature and pressure is not significant during the course of the run. Error bars on the structural and TPE distributions were found to be less than the symbol size in the associated figures.

3. RESULTS AND DISCUSSION

3.1. Tetrahedral Order and Energetics in Bulk Water. *3.1.1. Distributions of Tetrahedral Order, Tagged Potential Energy, and Coordination Number.* The liquid-state anomaly that is expected to affect solvation thermodynamics is the density anomaly, corresponding to a temperature–density regime for which $(\partial\rho/\partial T)_P > 0$. The boundary of this regime of anomalous density behavior is defined by state points for which $(\partial\rho/\partial T)_P = 0$, i.e., the locus of extrema of the $\rho(T)$ curve along isobars. In the case of water, the temperatures of minimum density occur in a thermodynamically unstable regime, and therefore only the locus of temperatures of maximum density (TMD) can be defined. The state point corresponding to the maximum temperature along the TMD locus in the ρ – T plane corresponds to the temperature of onset of thermodynamic anomalies and is denoted by $(T_{\text{TMD}}^{\text{max}}, \rho_{\text{TMD}}^{\text{max}})$. The onset temperature for thermodynamic anomalies of various rigid-body water models has recently been determined. The $(T_{\text{TMD}}^{\text{max}}, \rho_{\text{TMD}}^{\text{max}})$ state points for mTIP3P and TIP4P correspond to (202 K, 0.95 g cm⁻³) and (257 K, 0.98 g cm⁻³), respectively.⁴² The corresponding state point for the original TIP3P model is (205 K, 0.96 g cm⁻³).

Table 4. Average Pressure and Average Temperature Obtained from NVE Simulations for Solvated System of Peptide with mTIP3P and TIP4P

water model	target temperature (K)	$\langle T \rangle$ (K)	$\langle P \rangle$ (katm)	error _T (K)	error _P (katm)
mTIP3P	300	301.56	0.092	0.0102	0.0010
	250	252.09	0.249	0.0276	0.0033
TIP4P	300	305.05	-0.141	0.0345	0.0032
	250	253.09	0.097	0.0287	0.0031

Tetrahedral order distributions for water models in the anomalous regime are known to have a very prominent shoulder structure indicative of comparable fractions of tetrahedral and nontetrahedral local environments. In this structurally anomalous regime, tetrahedral order (q_{tet}) is strongly correlated with translational order (τ), which measures the extent of pair correlation between oxygen atoms.¹¹ The structurally anomalous regimes of all of the rigid-body water models are essentially superimposable in the q_{tet} – τ plane, even though the temperature and density regimes for the anomalies may be very different.⁴² An interesting illustration of this is shown in Figure 1a where the $P(q_{\text{tet}})$'s of mTIP3P, TIP4P, SPC/E, and TIP3P at their respective $T_{\text{TMD}}^{\text{max}}$ values are shown to be essentially superimposable. This was earlier shown to be true for the temperature of maximum density along the 1 atm isobar for some water models but not placed in the context of the similarity in the order maps.^{72,73} In the case of the two specific models of interest in this study (mTIP3P and TIP4P), the q_{tet} distributions of the two water models are very different at a given temperature; as illustrated in Figure 1b at (260 K, 1.00 g cm⁻³). The peak at 0.8–0.9 corresponding to strongly tetrahedral local environments is significantly diminished for the mTIP3P model indicating that mTIP3P behaves essentially as a simple liquid while TIP4P is an anomalous, tetrahedral liquid at 260 K.

To contrast the structural information contained in the tetrahedral order distributions with that in the radial distribution function, Figure 2b compares the oxygen–oxygen RDFs of TIP4P and mTIP3P at (260 K, 1.00 g cm⁻³). While the reduced structuring of the mTIP3P RDF compared to that of TIP4P is obvious, the nature of change in local order is much more evident from the $P(q_{\text{tet}})$ distributions in Figure 1b. Figure 2a shows that the RDFs of the two water models are identical at $(T_{\text{TMD}}^{\text{max}}, \rho_{\text{TMD}}^{\text{max}})$ as expected from the strong correlation between pair and tetrahedral ordering in the anomalous regime.

The distribution of tagged potential energies, $P(u_{\text{tag}})$, in rigid-body water models is approximately Gaussian, with a small asymmetric tail.^{51,74,75} Unlike the tetrahedral order distributions, there is no prominent shoulder in the anomalous regime, as shown in Figure 7a for the $P(u_{\text{tag}})$ distribution for TIP4P water at 260 K and 1.00 g cm⁻³.

To understand the contributions to u_{tag} from water molecules at different distances, we also compute the contribution to the tagged potential energy of a water molecule from water molecules restricted to lie in the first and second neighbor shells, denoted by u_{first} and u_{sec} , respectively. The cutoff distances for the first and second neighbor shells are taken as 3.5 Å and 5.5 Å, corresponding to the first and second minima in the $g_{\text{OO}}(r)$ pair correlation function, respectively, for TIP4P (see Figure 2b). Figure 3 shows the corresponding $P(u_{\text{first}})$ and $P(u_{\text{sec}})$ distributions. u_{first} is computed for a water molecule by summing the

Table 5. Comparison of Average Temperature and Average Pressure for the First 200 ps and Last 200 ps of the NVE Simulation of Peptide System Solvated with mTIP3P and TIP4P

water model	target temperature (K)	$\langle T_{\text{first}} \rangle$ (K)	$\langle T_{\text{last}} \rangle$ (K)	$\langle P_{\text{first}} \rangle$ (katm)	$\langle P_{\text{last}} \rangle$ (katm)
mTIP3P	300	301.48	301.86	0.0994	0.0844
	250	252.04	252.19	0.2487	0.2574
TIP4P	300	304.64	305.29	-0.1399	-0.1480
	250	252.91	253.72	0.0715	0.0837

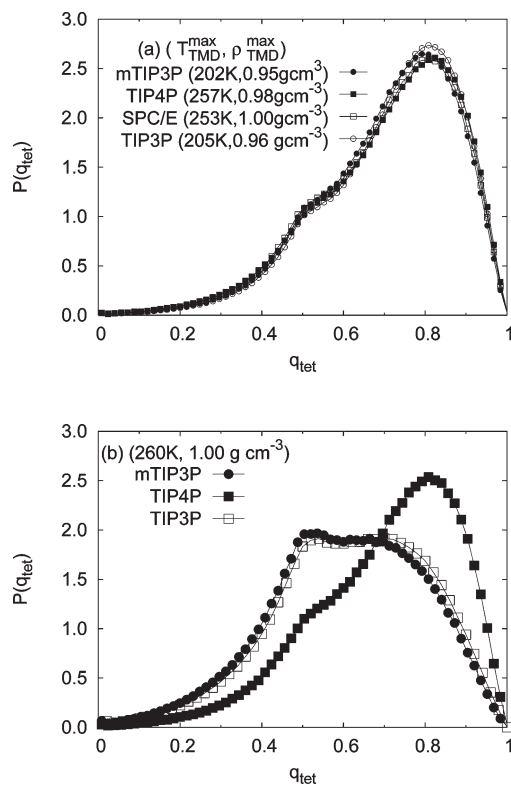


Figure 1. Normalized distributions of the tetrahedral order parameter (a) at the state point ($T_{\text{TMD}}^{\text{max}}, \rho_{\text{TMD}}^{\text{max}}$) corresponding to the maximum in temperature along the TMD locus for mTIP3P, TIP4P, SPC/E, and TIP3P water and (b) for mTIP3P, TIP4P, and TIP3P water at a common state point of (260 K, 1.00 g cm⁻³).

interaction energy of that molecule with waters lying within the first hydration shell, i.e., within 3.5 Å for TIP4P at (260 K, 1.00 g cm⁻³). Similarly, u_{sec} is computed for a water molecule by summing the interaction energy of that molecule with waters lying in the second shell, i.e., lying at a distance of 3.5–5.5 Å from that molecule. $u_{\text{first+sec}}$ is computed for a water molecule by summing the interaction energy of that molecule with all of the neighboring water molecules lying in both the first as well as the second shell of hydration, i.e., within 5.5 Å of that water molecule. So, $P(u_{\text{first+sec}})$ is the probability distribution of the $u_{\text{first+sec}}$ computed as above. Both the $P(u_{\text{first}})$ and $P(u_{\text{sec}})$ distributions are unimodal and do not show an obvious analogue of the strongly shouldered $P(q_{\text{tet}})$ distribution. The ensemble average of the combined contribution of the first and second neighbor shell corresponds to 92% of the ensemble-averaged TPE for TIP4P at

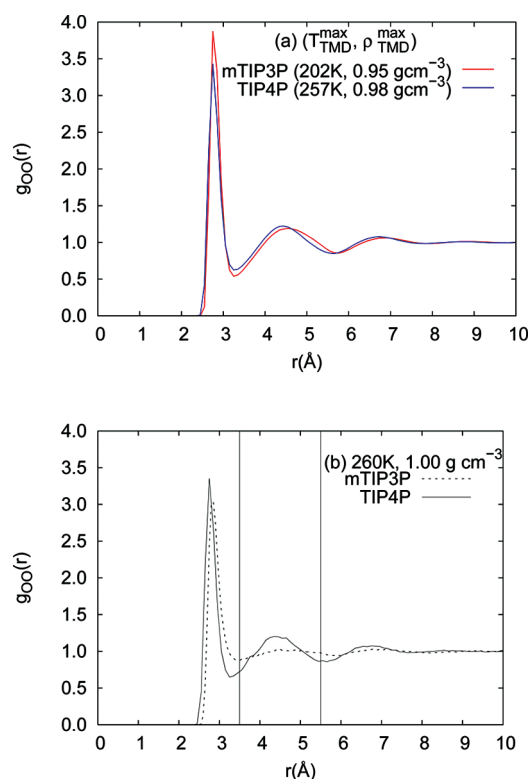


Figure 2. Radial distribution function of oxygen–oxygen, $g_{\text{OO}}(r)$, for mTIP3P and TIP4P water models at (a) their respective ($T_{\text{TMD}}^{\text{max}}, \rho_{\text{TMD}}^{\text{max}}$) state points and at (b) the state point (260 K, 1.00 g cm⁻³). The positions of vertical lines on the x axis indicate the upper limit of the first and the second hydration shells from left to right, respectively.

(260 K, 1.00 g cm⁻³), i.e., $\langle u_{\text{first+sec}} \rangle \approx \langle u_{\text{tag}} \rangle$ with molecules beyond the second shell contributing less than 10% to $\langle u_{\text{tag}} \rangle$. The ensemble-averaged value of TPE for TIP4P at (260 K, 1.00 g cm⁻³) is -87.031 kJ mol⁻¹.

The coordination number (n) of a given water is the number of neighbors lying within 3.5 Å, ensuring that the next nearest neighbor lies beyond 3.5 Å.^{76,77} The distribution of nearest neighbors in TIP4P and mTIP3P water at 260 K and 1.00 g cm⁻³ is shown in Figure 4a and b, respectively. In the case of TIP4P water at this state point, four- and five-coordination are almost equally probable, whereas in mTIP3P water, five-coordinate water molecules dominate. The average coordination number in TIP4P and mTIP3P water at this state point is 4.7 and 4.9, respectively. Neutron scattering data suggest that at 298 K and 1.00 g cm⁻³ each water molecule has ~4.5 neighbors.⁷⁸

3.1.2. Relationship between Local Order, Energy, and Coordination. It is evident that local order (q_{tet}), coordination number (n), and tagged molecule energy (u_{tag}) must carry correlated, but not identical, information about the local environment of water molecules. A simple correlation plot of the $q_{\text{tet},i}$ and $u_{\text{tag},i}$ for an individual molecule i sampled at a given state point, however, shows only very weak correlations between local order and energy. This lack of a well-defined relationship between the three quantities is also evident from the qualitatively different shapes of the shouldered $P(q_{\text{tet}})$ distributions compared with the unimodal $P(u_{\text{tag}})$ (Figure 7a) and $P(n)$ distributions (Figure 4).

Nonetheless, suitably defined averages of these quantities do show a good correlation. For example, the ensemble-averaged values of u_{tag} and q_{tet} along isochores have been shown to have a

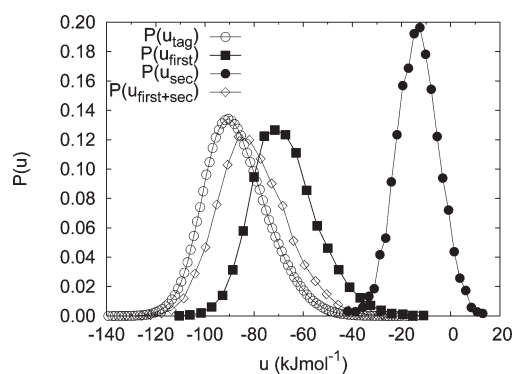


Figure 3. Normalized distributions of local energies of TIP4P water molecules at (260 K, 1.00 g cm⁻³). The local energy is calculated as a sum of contributions from all other $N - 1$ molecules in the system and is denoted by u_{tag} . u_{first} and u_{sec} refer to local energies due to interactions with first and second shell neighbors, respectively; $u_{\text{first+sec}}$ refers to the combined contribution to local energy from the first and second shell neighbors.

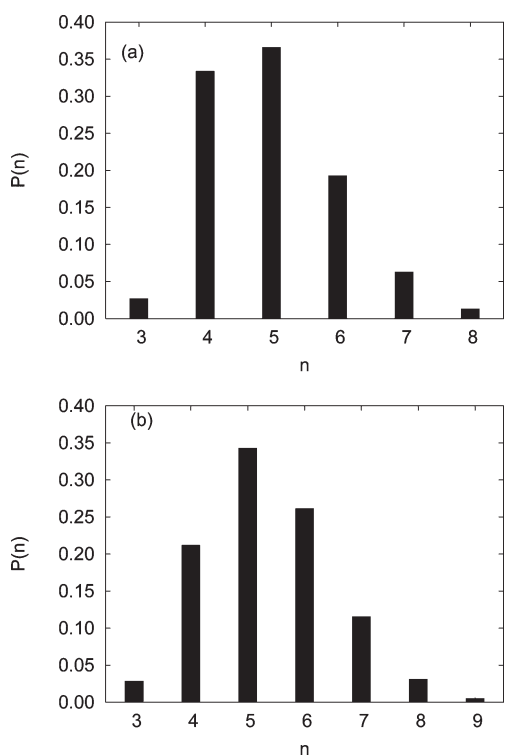


Figure 4. Coordination number distributions, $P(n)$, computed at (260 K, 1.00 g cm⁻³) for (a) TIP4P and (b) mTIP3P. The probability of finding a higher coordination number ($n > 6$) of water is very low. $\sum_{n=1}^8 P_n = 0.998$ for TIP4P and $\sum_{n=1}^9 P_n = 0.997$ for mTIP3P.

well-defined linear correlation in previous studies.⁵¹ Similarly, the tagged particle potential energy, $\langle u_{\text{tag}} \rangle_{q_{\text{tet}}}$ obtained by averaging over water molecules with local tetrahedral order lying between q_{tet} and $q_{\text{tet}} + \delta q_{\text{tet}}$ shows a quasi-linear, negative correlation with q_{tet} .⁷³ In Figure 5, we show these correlations for selected state points of TIP4P water.

It is useful to employ other metrics in addition to correlation plots to characterize the relationships between local order, energy, and coordination number. We have found it convenient

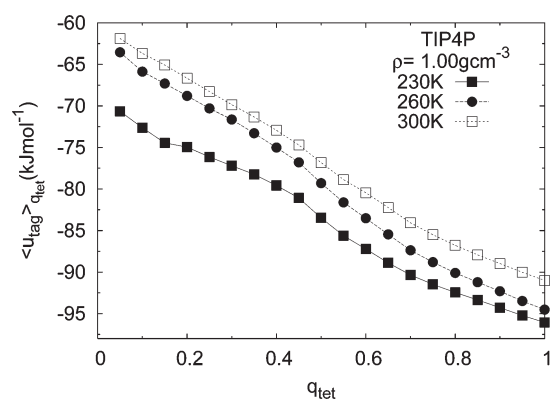


Figure 5. Correlations between local order and local energy averages for TIP4P water molecules at 1.00 g cm⁻³ and temperatures of 230, 260, and 300 K. The average local energy $\langle u_{\text{tag}} \rangle_{q_{\text{tet}}}$ is calculated by averaging u_{tag} over molecules having q_{tet} values lying between q_{tet} and $q_{\text{tet}} + \delta q_{\text{tet}}$ ($\delta q_{\text{tet}} = 0.05$).

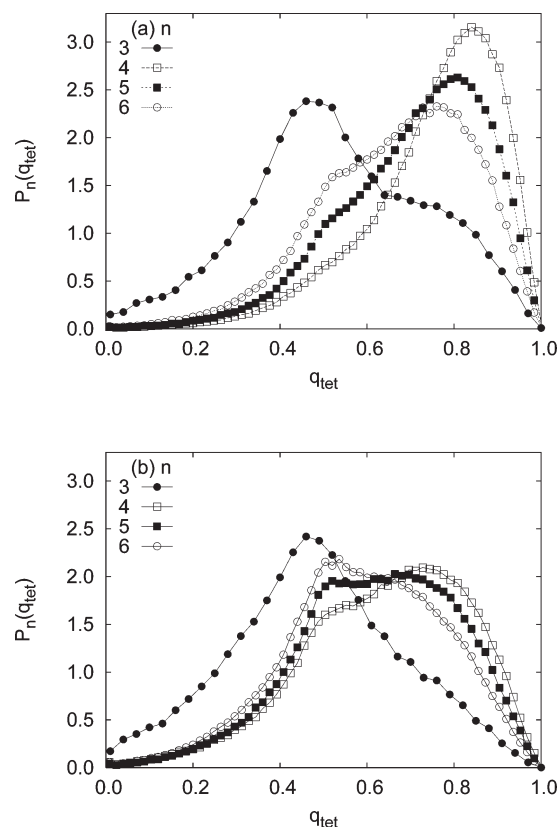


Figure 6. Normalized conditional probability distribution of local order, $P_n(q_{\text{tet}})$ for water molecules with coordination number (n) computed at (260 K, 1.00 g cm⁻³) for (a) TIP4P and (b) mTIP3P.

to use conditional distributions of one quantity subject to a restricted range of another quantity. For example, one can create distributions of the tetrahedral order and the TPE using only the subset of water molecules with coordination number n , which will be denoted by $P_n(q_{\text{tet}})$ and $P_n(u_{\text{tag}})$. Similarly one can consider the distributions, $P_u(q_{\text{tet}})$, corresponding to the distributions of q_{tet} for all molecules with tagged potential energy lying between u and $u + \delta u$ where u is very small compared with the range of accessible values of u .

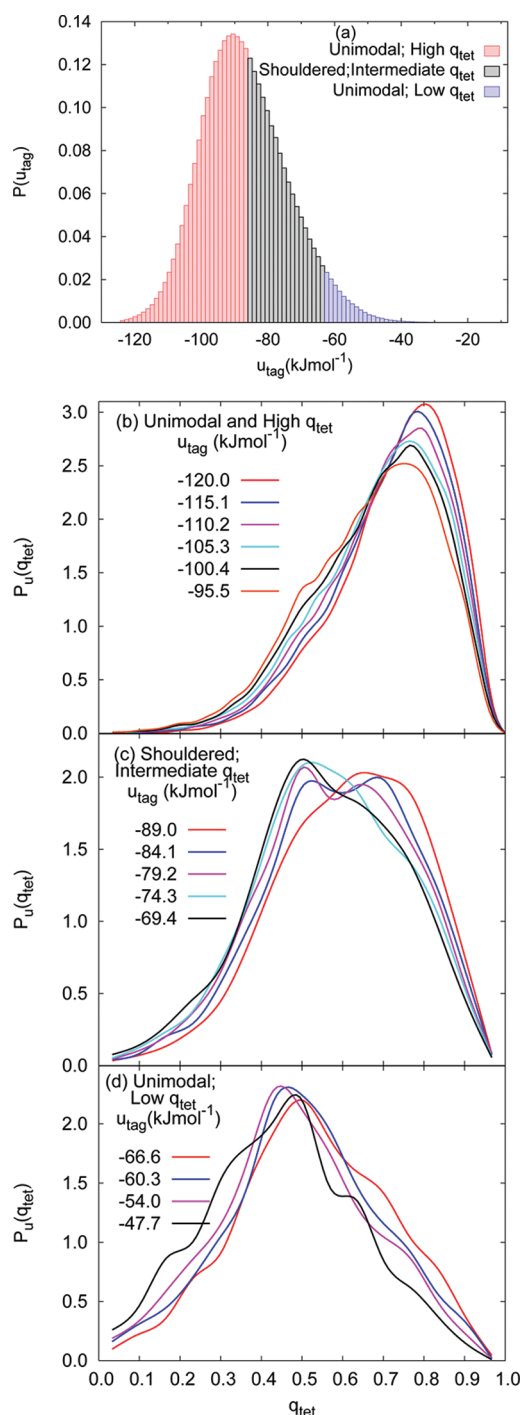


Figure 7. Relationship between tagged molecule potential energy distribution, $P(u_{\text{tag}})$, and tetrahedral order distribution, $P_u(q_{\text{tet}})$, conditional on water molecules having u_{tag} lying between u and $u + du$. Results are shown for TIP4P water at (260 K, 1.00 g cm⁻³). Part a shows a normalized $P(u_{\text{tag}})$ distribution ($\delta u = 0.7$ kJ mol⁻¹) colored according to three types of normalized $P_u(q_{\text{tet}})$ distributions shown in parts b, c, and d. Part b shows unimodal $P_u(q_{\text{tet}})$ distributions for water molecules with high q_{tet} having $-120 \leq u_{\text{tag}} \leq -90$ kJ mol⁻¹. Part c shows strongly shouldered $P_u(q_{\text{tet}})$ distributions for molecules with intermediate q_{tet} having $-90 \leq u_{\text{tag}} \leq -65$ kJ mol⁻¹. Part d shows unimodal $P_u(q_{\text{tet}})$ distributions for water with a nontetrahedral environment having $-65 \leq u_{\text{tag}} \leq -40$ kJ mol⁻¹. The red, black, and blue regions of $P(u_{\text{tag}})$ distribution correspond to water molecules having $P_u(q_{\text{tet}})$ distributions shown in parts b, c, and d, respectively.

Figure 6 shows the conditional tetrahedral order distributions, $P_n(q_{\text{tet}})$, for TIP4P and mTIP3P water at 260 K and 1.00 g cm⁻³. In the case of TIP4P water, it is immediately obvious that the four-coordinate molecules are preferentially in a strongly tetrahedral environment. The $P_{n=4}(q_{\text{tet}})$ distribution is unimodal with a sharp peak at $q_{\text{tet}} \approx 0.8$. The five- and six-coordinate water molecules show shouldered $P(q_{\text{tet}})$ distributions with a shoulder at $q_{\text{tet}} \approx 0.45$. The three-coordinate water molecules, which form only a 0.029 fraction of water molecules, are far from tetrahedrality having unimodal $P(q_{\text{tet}})$ distribution with a peak at $q_{\text{tet}} \approx 0.4$. In the case of mTIP3P water, the $P_{n=3}(q_{\text{tet}})$ distribution is very similar to that of TIP4P. The conditional tetrahedral order distributions for $n = 4, 5$, and 6 are very similar, shouldered distributions with peaks at $q_{\text{tet}} \approx 0.5$ and $q_{\text{tet}} \approx 0.7$ of almost equal height.

Figure 7a shows the $P(u_{\text{tag}})$ distribution of TIP4P water at (260 K, 1.00 g cm⁻³) subdivided into histogram bins of approximately 1 kJ mol⁻¹ width. The $P_u(q_{\text{tet}})$ distributions in each bin were generated and could be classified into three types (Figure 7b–d). For the low TPE histograms, lying in the range of -120 to -90 kJ mol⁻¹, the $P_u(q_{\text{tet}})$ distributions are unimodal with a dominant peak corresponding to a strongly tetrahedral local environment at $q_{\text{tet}} \approx 0.85$ and a weak shoulder at approximately $q_{\text{tet}} \approx 0.5$. Comparison with Figure 6a suggests that water molecules in this energy region are four-coordinated. Water molecules with u_{tag} values in the range of -90 to -65 kJ mol⁻¹ have local environments of intermediate tetrahedrality. $P_u(q_{\text{tet}})$ distributions for these water molecules have a peak at $q_{\text{tet}} \approx 0.75$ and a second one at $q_{\text{tet}} \approx 0.45$. Both peaks are of comparable height, with the $q_{\text{tet}} \approx 0.45$ peak becoming more prominent with increasing u_{tag} . The water molecules in the high energy tail of the u_{tag} distribution (-65 to -40 kJ mol⁻¹) have nontetrahedral local environment with unimodal $P_u(q_{\text{tet}})$ distributions having a peak lying below ~ 0.43 . Thus, it is possible to color the histogram bins in the $P(u)$ distributions in terms of three qualitatively different types of $P_u(q_{\text{tet}})$ distributions: (i) strongly tetrahedral with a sharp peak for q_{tet} lying between approximately 0.8 and 0.9, (ii) intermediate tetrahedral character with significant proportion of tetrahedral and nontetrahedral environments, and (iii) nontetrahedral distributions with a peak at q_{tet} of about 0.5. The tetrahedral order distributions conditional on a particular value of the tagged particle energy show that TIP4P water molecules with low local energies will exist preferentially in tetrahedral environments, illustrating the energetic bias for tetrahedrality for this model at the given state point.

Figure 8 shows the conditional $P_u(q_{\text{tet}})$ for mTIP3P water at 260 K and 1.00 g cm⁻³. In striking contrast to the results for TIP4P, the $P_u(q_{\text{tet}})$'s are essentially identical for the entire range of u_{tag} values and consist of distributions indicating comparable proportions of tetrahedral and nontetrahedral local environments. In the case of mTIP3P, there is no energetic bias for tetrahedrality at a temperature of 260 K.

The shouldered distributions of $P(q_{\text{tet}})$ in the anomalous regime of the water models show that the free energy of the system as a function of local order must have two local minima. Since the local energy distributions do not show a corresponding presence of tetrahedral and nontetrahedral environments, one expects that entropic factors must play an important role. In a recent work, Stanley and co-workers introduced an entropy measure based on the tetrahedral order distribution.⁷⁹ We adapt it to define an entropy measure, $S_q(u_{\text{tag}})$, using the conditional $P_u(q_{\text{tet}})$ distribution for all water molecules with u_{tag} lying

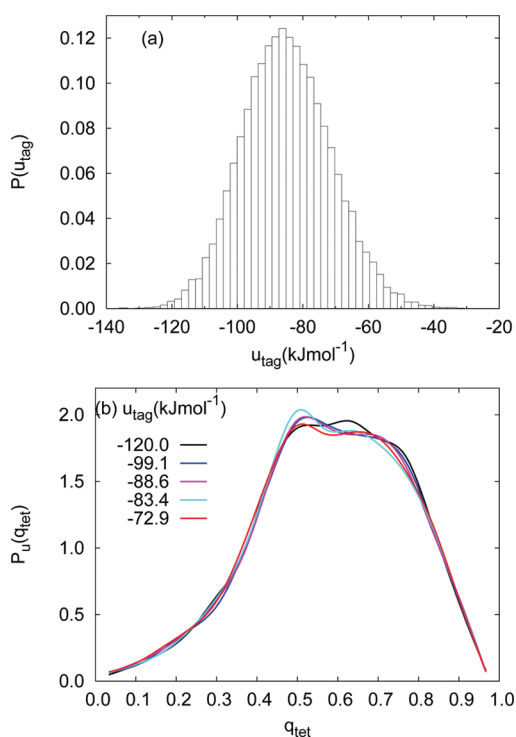


Figure 8. Relationship between tagged molecule potential energy distribution, $P(u_{\text{tag}})$, and tetrahedral order distribution, $P_u(q_{\text{tet}})$, conditional on water molecules having u_{tag} lying between u and $u + \delta u$. Results are shown for mTIP3P water at (260 K, 1.00 g cm^{-3}). Part a shows normalized $P(u_{\text{tag}})$ distribution ($\delta u = 1.05 \text{ kJ mol}^{-1}$). Part b shows $P_u(q_{\text{tet}})$ distributions for water molecules with u_{tag} lying between u and $u + \delta u$ for the entire range of u_{tag} shown in part a. mTIP3P does not show different types of $P_u(q_{\text{tet}})$ distributions for different ranges of u_{tag} at this state point as are shown by TIP4P (Figure 7b–d).

between u and $u + \delta u$ at a given state point.

$$S_q(u_{\text{tag}}) = S_0 + \frac{3}{2} k_B \int_{q_{\text{tet}}^{\min}}^{q_{\text{tet}}^{\max}} \ln(1 - q_{\text{tet}}) P_u(q_{\text{tet}}) dq_{\text{tet}} \quad (3)$$

$S_q(u_{\text{tag}})$ is measured relative to S_0 where $S_0 = k_B [\ln \Omega_0 + 3/2 \ln(8/3)]$ at all state points.⁷⁹ The q_{tet}^{\min} and q_{tet}^{\max} correspond to the minimum and maximum values of q_{tet} of water molecules having u_{tag} in the range of u and $u + \delta u$. $S_q(u_{\text{tag}})$ is calculated in different ranges of u_{tag} described above, and its behavior is examined as a function of u_{tag} for mTIP3P and TIP4P (see Figure 9).

It is observed that for TIP4P at (260 K, 1.00 g cm^{-3}), $S_q(u_{\text{tag}})$ shows a positive correlation with u_{tag} and anticorrelation with q_{tet} , indicating that the higher the q_{tet} the more negative is the tetrahedral entropy, and the lower the u_{tag} the lower is the tetrahedral entropy of the system. At 350 K and 1.00 g cm^{-3} for TIP4P water, the correlation is somewhat weaker. For mTIP3P, there is no correlation between $S_q(u_{\text{tag}})$ and u_{tag} at 260 K. At the maximum temperature along the TMD locus, the $S_q(u_{\text{tag}})$ plots for mTIP3P and TIP4P are very similar.

3.2. Tetrahedral Order and Local Energetics of Water in Hydration Shells. Figure 10 compares the average number of water molecules $n(r)$ present at a distance between r and $r + \delta r$ ($\delta r = 0.25 \text{ \AA}$) from the β -hairpin peptide solvated in mTIP3P and TIP4P water at 250 and 300 K. The distance of the water molecule was measured relative to the nearest atom of the

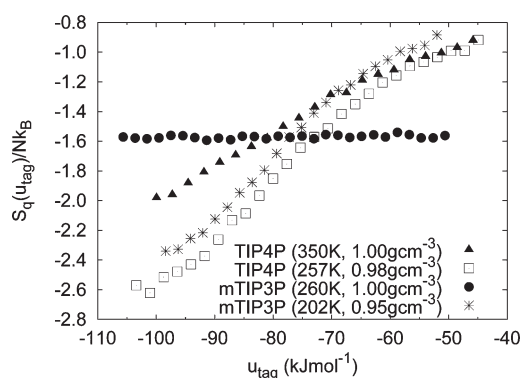


Figure 9. Plot of tetrahedral entropy, $S_q(u_{\text{tag}})$, defined in eq 3 as a function of u_{tag} for mTIP3P and TIP4P at various state points. (T_{TMD}^{\max} , ρ_{TMD}^{\max}) for mTIP3P is (202 K, 0.95 g cm^{-3}) and for TIP4P is (257 K, 0.98 g cm^{-3}).

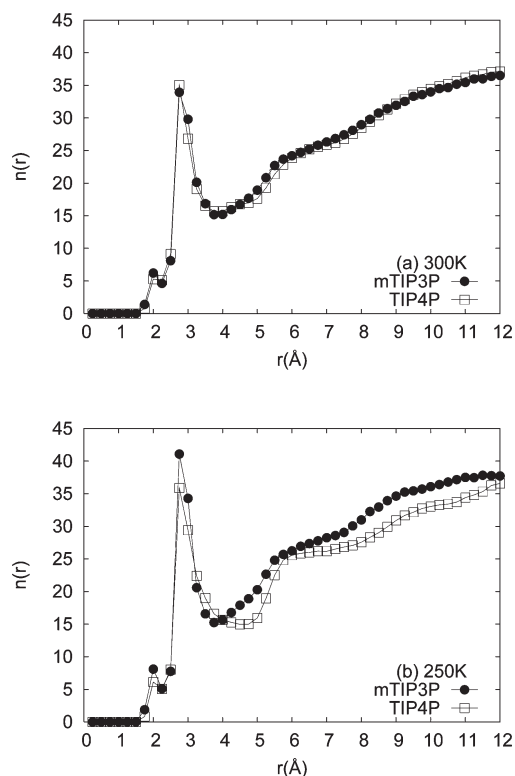


Figure 10. Comparison of the number of water molecules, $n(r)$, as a function of distance r from 2GB1 β -hairpin peptide when solvated in mTIP3P and TIP4P water at (a) 300 K and (b) 250 K.

peptide. At 300 K, $n(r)$ does not change significantly with the change in water models, mTIP3P and TIP4P, in agreement with previous studies.⁵⁰ At 250 K, small differences between the solvent models can be seen in the neighborhood of the first hydration shell, with a sharper peak coming from the mTIP3P model. A weak layering effect which extends up to 12 \AA from the peptide is seen for both the water models.⁸⁰

Figure 11 shows the radial variation of the tagged potential energy as a function of the distance from the β -hairpin peptide at 300 and 250 K for both water models. The $u_{\text{tag}}(r)$ curves are generated by calculating the mean u_{tag} values of all of the $n(r)$ water molecules lying at a distance between r and $r + \delta r$ from the

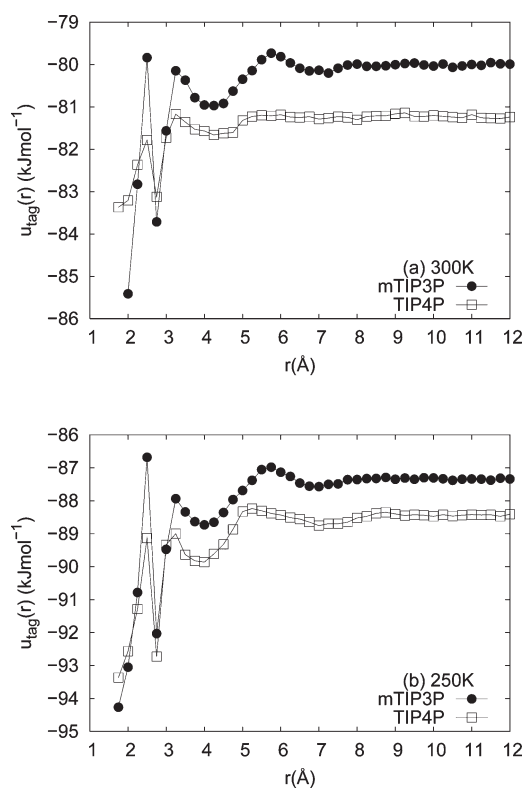


Figure 11. Comparison of tagged potential energy of water molecules, $u_{\text{tag}}(r)$, as a function of the distance (r) from the 2GB1 β -hairpin peptide when solvated in mTIP3P and TIP4P water at (a) 300 K and (b) 250 K. The value of $u_{\text{tag}}(r)$ at 12 Å is close to the bulk $\langle u_{\text{tag}} \rangle$ value at 1.00 g cm⁻³. The values of $\langle u_{\text{tag}} \rangle$ at 1.00 g cm⁻³ are -81.88 kJ mol⁻¹ and -81.29 kJ mol⁻¹ for TIP4P and mTIP3P at 300 K, respectively, and -88.41 kJ mol⁻¹ and -86.30 kJ mol⁻¹ for TIP4P and mTIP3P at 250 K, respectively.

peptide. Our previous study shows an oscillatory variation in $u(r)$ beyond the first hydration shell in mTIP3P water.⁵¹ The oscillatory behavior of TIP4P water is very similar to that of mTIP3P solvent, except that the oscillations are less pronounced in TIP4P. The $u_{\text{tag}}(r)$ values are lower for the TIP4P model by approximately 1 kJ mol⁻¹ at both 300 and 250 K.

Figure 12 shows the tetrahedral order parameter as a function of the distance from the peptide. As discussed in our previous study, beyond the first hydration shell, the maxima and minima in the $q_{\text{tet}}(r)$ distribution tend to be anticorrelated with those in $u_{\text{tag}}(r)$.⁵¹ As expected on the basis of bulk behavior, TIP4P solvent shows significantly higher local tetrahedral order compared with the mTIP3P solvent in the hydration shell of a peptide.

The results presented in Figures 10–12 focus on the local energy and structure beyond the first hydration shell. Within the first hydration shell, we expect the tetrahedral network of water to be distorted, in addition to a strong dependence on the chemical character of the residues.

To characterize the differences between the two water models within this critical solvent layer, Figure 13 shows mean TPEs of water molecules lying within 5 Å of any atom of each amino acid residue. The variation in TPE as a function of the residues of the peptide shows similar qualitative behavior for both mTIP3P and TIP4P water models; i.e., the water molecules near hydrophilic residues 42(Glu), 46(Asp), 47(Asp), and 56(Glu) have lower TPEs and those near hydrophobic residues 43(Trp) and 52-(Phe) have higher TPE values at 300 K. The quantitative

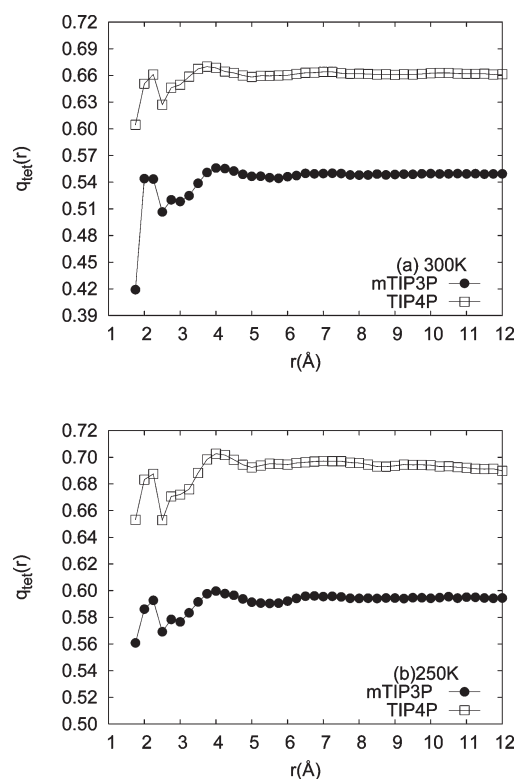


Figure 12. Comparison of tetrahedral order, $q_{\text{tet}}(r)$, as a function of distance (r) from the 2GB1 β -hairpin peptide when solvated in mTIP3P and TIP4P water at (a) 300 K and (b) 250 K.

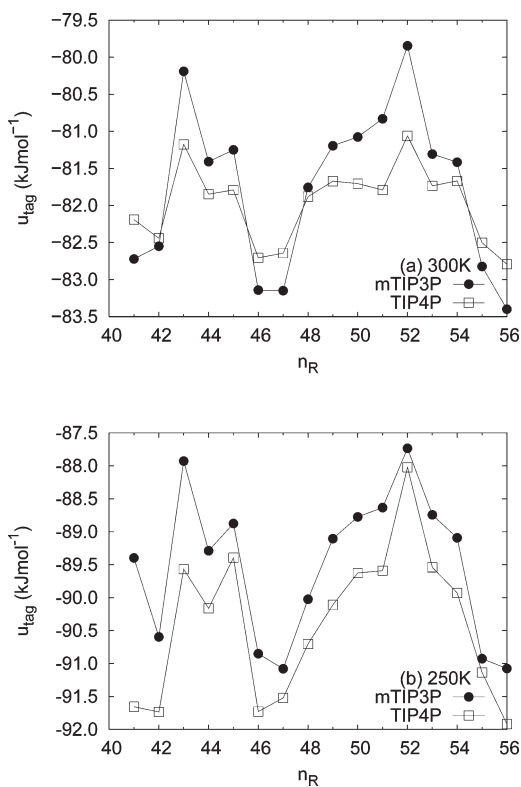


Figure 13. Residue-wise dependence of tagged potential energy for mTIP3P and TIP4P at (a) 300 K and (b) 250 K.

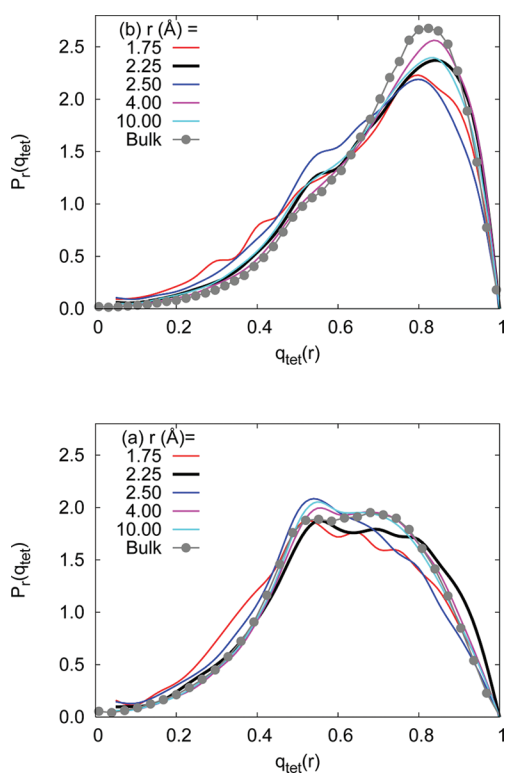


Figure 14. Normalized distributions of tetrahedral order, $P_r(q_{\text{tet}})$, for water molecules lying between a distance of r and $r + \delta r$ ($\delta r = 0.25 \text{ \AA}$) from the peptide solvated in (a) mTIP3P and (b) TIP4P at 250 K. The bulk $P(q_{\text{tet}})$ distributions shown correspond to the state point (250 K, 1.00 g cm^{-3}) for both the water models.

differences in TPE of the water molecules near the hydrophilic and hydrophobic residues are, however, greater in the case of mTIP3P than in TIP4P.

Figure 14 shows the $P_r(q_{\text{tet}})$ distributions for water molecules lying between a distance of r and $r + \delta r$ of the peptide solvated in mTIP3P as well as TIP4P water at 250 K. The $P_r(q_{\text{tet}})$'s are compared with the bulk $P(q_{\text{tet}})$ distributions of the two water models. The distances corresponding to minima in the $q_{\text{tet}}(r)$ distributions shown in Figure 12 show tetrahedral order distributions with somewhat higher probabilities of low q_{tet} values. It is evident that deviations from the bulk distribution are small as a function of the distance from the peptide. It should be noted that q_{tet} has been defined to include the four nearest heavy atoms, whether they belong to water or the peptide. Since the oxygen atom in water will tend to form two donor and two acceptor hydrogen bonds, this definition of q_{tet} assumes that replacement of water–water hydrogen bonds with other hydrogen bonds (e.g., water and hydroxyl group of side chain) does not significantly alter the local environment of a water molecule. The distortion in the tetrahedral network of the water molecules in the first hydration shell is not as severe as expected, even though the TPE values of water molecules in the first hydration shell are significantly lower. It should be noted that similar small perturbations of the tetrahedral order distribution as a function of the radial distance from the solute were seen in the case of aqueous sugar solutions.⁸¹

4. CONCLUSIONS

We show that the relationship between tetrahedral order and anomalous behavior is very similar in the rigid-body water

models (TIP3P, mTIP3P, SPC/E, and TIP4P); for example, the tetrahedral order distributions for mTIP3P, TIP4P, and SPC/E water are superimposable at the state point ($T_{\text{TMD}}^{\text{max}}$, $\rho_{\text{TMD}}^{\text{max}}$) corresponding to the maximum temperature for onset of the density anomaly. The energetic stabilization of tetrahedrality for the different water models can vary widely, as indicated by the wide variation in the temperature regimes of the anomalies. In the temperature range of 250–300 K, of interest in biomolecular simulations, the mTIP3P water model behaves as a normal liquid, while the TIP4P water model is an anomalous, tetrahedral liquid. To understand the relationship between local order (q_{tet}), local energy (u_{tag}), and local coordination number (n) in these two water models, we construct tetrahedral order distributions conditional on a specific value of the coordination number or a narrow range of tagged molecule potential energies. The $P_n(q_{\text{tet}})$ distributions are obtained by averaging over all molecules with exactly n neighbors in the first coordination shell. The $P_u(q_{\text{tet}})$ distributions are obtained by averaging over all water molecules with TPE lying between u and $u + \delta u$. The $P_n(q_{\text{tet}})$ or $P_u(q_{\text{tet}})$ for TIP4P can be classified into three types: (i) strongly tetrahedral with a sharp peak for q_{tet} lying between 0.8 and 0.9, (ii) intermediate tetrahedral character with a broad plateau or strong shoulder indicating a significant proportion of both tetrahedral and nontetrahedral environments, and (iii) nontetrahedral distributions with a peak at q_{tet} of about 0.5. Using the conditional distributions, we show that four-coordinate environments in TIP4P water in the 250–270 K regime are likely to be strongly tetrahedral with lower local tagged molecule potential energies. Higher coordination sites will have lower tetrahedrality and higher local energy. The mTIP3P model of bulk water in the same temperature regime displays no energetic bias toward tetrahedral local environments. The tendency of water molecules with different binding energies (TPEs) to occupy environments with different degrees of disorder can be effectively illustrated by constructing entropy measures based on the $P_u(q_{\text{tet}})$ distributions. At the maximum temperature along the TMD locus, the correlation between tetrahedral entropy versus tagged potential energy is strong and virtually identical for mTIP3P and TIP4P. In TIP4P, this correlation is retained up to temperatures as high as 300 K, while it is lost by 250 K in mTIP3P.

We compare the effect of solvating the 16-residue β -hairpin fragment of 2GB1 in mTIP3P and TIP4P solvents. The hydration shell coordination profile ($n(r)$) of the number of water molecules at a distance r from the peptide is almost identical in the two solvents at 300 K, though lowering of the temperature to 250 K increases the differences slightly. This is consistent with earlier studies by Nutt and Smith on the radial distribution functions in water.⁵⁰ The differences between the two solvent models are larger, however, for the tagged potential energy ($u_{\text{tag}}(r)$) as a function of distance r from the peptide, though the weak oscillatory variation as a function of r is qualitatively very similar in the two solvents. The residue-wise variation in the tagged potential energy of water molecules within the first hydration shell is qualitatively similar in the two models. The differences between the two solvent models are larger when the radial distribution of the tetrahedral order is monitored, compared to the tagged potential energy. This difference, however, seems to originate from the differences in the bulk tetrahedral order distributions of the solvent at the same state point since perturbation in the local tetrahedral order distributions of the bulk solvent due to the presence of the solute is marginal.

To summarize, the mTIP3P and TIP4P water models show qualitatively different behavior in terms of the relationship between tetrahedral order and local energy that governs the thermodynamic anomalies in the 250–300 K temperature regime. As solvents in the neighborhood of a biomolecular solute, the differences between the two models, however, are quantitative rather than qualitative.

It is useful to place our results in the context of previous work on structure and energetics in bulk water and in aqueous solutions. We show that local order, energy, and coordination number in bulk water show similar but not identical behavior in the context of the liquid-state anomalies. This is consistent with the multiple time-scale behavior seen in the time correlation of local fluctuations in tetrahedral order, coordination number, and tagged particle potential energies.^{7,56,57,76,82} Our results for the differences in shapes of tetrahedral order distributions for water molecules with different local coordination are also consistent with the recent work of Netz and co-workers that density–density correlations do not have a simple relationship with spatial correlations in structural order.⁸³ Previous studies suggest that rigid-body water models that provide a more accurate phase diagram and equation of state, also prove to be more reliable for studying solvation and clathrate formation of small hydrophobes.^{26,27} A systematic comparison of hydration behavior of amino acid analogues, small solutes, and proteins (e.g., myoglobin and crambin) predicted by different biomolecular force fields and water models suggests that qualitative trends are similar, though the equation of state of the water model strongly influences the hydration enthalpy.^{31,49,50} These results are consistent with our observations on the hydration shell structure and energetics of the 2GB1 peptide in water.

The results of this study provide some insight into the role played by different components of the intermolecular interactions of water in controlling hydration behavior, which should help in the appropriate design of coarse-grained models for water.^{84–86} A crucial factor controlling the chemical potential of small hydrophobes in water is the cavity distribution function which depends on the size of the water molecules.^{27,28} In the TIPnP water models, this aspect will be controlled essentially by the Lennard-Jones size parameter and will be identical for all of them. The geometry of the partial charges is the crucial difference between the two models studied here. As pointed out by Vega and others, the dipole moments of most rigid-body water models are very similar, though the quadrupole moments are significantly different.^{40,61} The similar dipole moments of the two models studied here imply that the most important contribution to solvation of polar solutes or amino acid side chains is similar in the mTIP3P and TIP4P models. The dipole–quadrupole ratio has been shown to be crucial in determining the phase diagram and equation of state of water, especially the regions of stability of the different ice phases.⁶¹ One may therefore surmise that this factor will affect organization of the tetrahedral network and therefore the presence of water-like anomalies, solvation of small hydrophobes, and clathrate formation.^{26,27} Given the length-scale dependence of hydrophobicity, the dipole–quadrupole ratio and the associated tetrahedral hydrogen-bonded network structure may play a relatively small role in the hydration of a peptide of nanoscale dimensions with a heterogeneous set of charged, polar, and nonpolar residues, as suggested by our results. The qualitatively similar behavior in local energy and order in the peptide hydration shell in mTIP3P and TIP4P water shown by us leads to the conclusion that the dipole moment and

the van der Waals radius play a crucial role in determining overall trends in hydration properties, while the dipole–quadrupole ratio and equation of state of water play a subsidiary though significant role.

APPENDIX A: EVALUATION OF TAGGED MOLECULE POTENTIAL ENERGY FOR PAIR-ADDITIVE POTENTIALS

In this section, we present the equations required in order to evaluate the tagged molecule potential energy (TPE or u), defined as the interaction of a given molecule with all other molecules in the system. Our equations apply to the case where intermolecular forces are modeled using pair interactions between Lennard-Jones and charge sites and do not take into account intramolecular contributions. The Coulombic interactions are evaluated using Ewald summation. Regarding computational details, we have computed the TPE of water molecules in rigid-body models of bulk water using these equations by adapting standard Ewald summation codes.⁸⁸ We have then shown that the TPE can be computed by relatively small modifications of the DL_POLY source code. The *pairInteraction* command in the NAMD input file was then used to also compute the TPE distributions for bulk water and was shown to give essentially identical results, though NAMD uses the particle-mesh Ewald method to evaluate the Coulombic forces. With this validation, we could use the *pairInteraction* command in NAMD to compute the TPEs of water molecules in the hydration shell of a biomolecule.

In the case of pair-additive potential models, such as the CHARMM biomolecular force field, the intermolecular contribution to the potential energy (U_{inter}) can be written as a sum of the Coulombic (U_{Coul}) and van der Waals (U_{vdW}) interactions.

$$U_{\text{inter}} = U_{\text{Coul}} + U_{\text{vdW}} \quad (\text{A.1})$$

In the Ewald summation, U_{Coul} is written as

$$U_{\text{Coul}} = U_{\text{screen}} + U_{\text{rec}} - U_{\text{self}} \quad (\text{A.2})$$

where U_{screen} is the real space contribution from screened charges, U_{rec} is the reciprocal space contribution, and U_{self} is the self-interaction correction.^{64,87,88} The tagged molecule potential energy (TPE) of the i th molecule will therefore contain a corresponding set of contributions and will be written as

$$u_i = u_i^{\text{screen}} + u_i^{\text{rec}} - u_i^{\text{self}} + u_i^{\text{vdW}} \quad (\text{A.3})$$

In the case of the one-component bulk water system, the tagged molecule energies will satisfy the relation

$$U_{\text{inter}} = \frac{1}{2} \sum_i u_i \quad (\text{A.4})$$

The rigid-body models associate a set of Lennard-Jones (LJ) and charged sites with a water molecule.

The TPE contribution due to van der Waals interactions will therefore be

$$u_i^{\text{vdW}} = \sum_k^{N_{\text{LJ}}} u_k^{\text{vdW}} = \sum_k \sum_j^{N_{\text{LJ}}} u_{\text{LJ}}(r_{kj}) \quad (\text{A.5})$$

where the sum over k represents summation of over all N_{LJ} Lennard-Jones sites on molecule i and the sum over j is the interaction of LJ site k with all LJ sites not on molecule i .

In the case of Ewald summation, it is necessary to consider the interaction of each of the charge sites l with all other charge sites not on the same molecule. Ewald summation replaces the bare charges by screened Gaussian charge distributions to give a real-space contribution to the potential energy, written as

$$U_{\text{screen}} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0} \frac{\text{erfc}(\alpha|r_{ij}|)}{|r_{ij}|} \quad (\text{A.6})$$

$$= \frac{1}{2} \sum_{j=1}^N u_j^{\text{creal}} \quad (\text{A.7})$$

where u_j^{creal} is the interaction of a screened charge site with all other screened charges. The u_i^{screen} contribution to the TPE of the i th molecule must be computed by summing the u_j^{creal} contributions due to all N_C charged sites on molecule i .

The reciprocal-space contribution to the Coulombic energy, U_{rec} is written as

$$U_{\text{rec}} = \frac{1}{2V_0\epsilon_0} \sum_{|k| \neq 0} \frac{\exp(-k^2/4\alpha^2)}{k^2} \left| \sum_j q_j \exp(-ik \cdot \mathbf{r}_j) \right|^2 \quad (\text{A.8})$$

where j labels the charge sites in the system. By replacing the squared modulus term by a double sum over i and j , we can write

$$U_{\text{rec}} = \frac{1}{2V_0\epsilon_0} \sum_{|k| \neq 0} \frac{\exp(-k^2/4\alpha^2)}{k^2} \sum_{j=1}^N \sum_{i=1}^N q_i q_j \exp(-ik \cdot (\mathbf{r}_j - \mathbf{r}_i)) \quad (\text{A.9})$$

By rearranging, one can show that it is possible to write U_{rec} as a sum of contributions from each charge site j of the form

$$U_{\text{rec}} = \frac{1}{2} \sum_{j=1}^N u_j^{\text{crec}} \quad (\text{A.10})$$

where u_j^{crec} is the reciprocal space contribution to the energy of interaction of a charge site with all other charges in the system and may be explicitly written as

$$u_j^{\text{crec}} = \frac{1}{V_0\epsilon_0} \sum_{|k| \neq 0} q_j \exp(-ik \cdot \mathbf{r}_j) \frac{\exp(-k^2/4\alpha^2)}{k^2} \sum_{i=1}^N q_i \exp(ik \cdot \mathbf{r}_i) \quad (\text{A.11})$$

The u_i^{rec} contribution to the TPE of the i th molecule must be computed by summing the u_j^{crec} contributions due to all N_C charged sites on molecule i .

The self-interaction term U_{self} corrects for the spurious interaction energy in U_{rec} of each screened charge distribution with itself as well as removes the interactions between charged sites on the same molecules. For a molecule i , this term can be written as

$$u_i^{\text{self}} = \frac{1}{4\pi\epsilon_0} \sum_{l \leq m}^{N_C} \left(q_l q_m \delta_{lm} \frac{\alpha}{\sqrt{\pi}} + \frac{\text{erf}(\alpha r_{lm})}{r_{lm}^{1-\delta_{lm}}} \right) \quad (\text{A.12})$$

where l and m label charge sites on the same molecule i and N_C is the number of charge-carrying sites per molecule.

AUTHOR INFORMATION

Corresponding Author

*Tel.: (+)91-11-2659-1510. Fax: (+)91-11-2686-2122. E-mail: charus@chemistry.iitd.ernet.in.

ACKNOWLEDGMENT

This work was financially supported by the Department of Science and Technology, New Delhi. D.N. would like to thank the Council of Scientific and Industrial Research for the award of a Junior Research Fellowship. M.A. would like to thank the Indian Institute of Technology—Delhi for the award of a Senior Research Fellowship. C.C. and D.N. would like to thank Sanjoy Bandhyopadhyay and Sudipta Kumar Sinha for useful discussions.

REFERENCES

- (1) Dill, K.; Truskett, T.; Vlachy, V.; Hribar-Lee, B. Modelling water, the hydrophobic effect and ion solvation. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 173–199.
- (2) Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **2005**, *437*, 640–647.
- (3) Ball, P. Water as an active constituent in cell biology. *Chem. Rev.* **2008**, *108*, 74–108.
- (4) Debenedetti, P. G. Supercooled and glassy water. *J. Phys., Cond. Matt.* **2003**, *15*, 1669–1726.
- (5) Bellissent-Funel, M.-C. *Hydration Processes in Biology: Theoretical and Experimental Approaches*; IOS Press: Amsterdam, The Netherlands, 1999.
- (6) Li, Z.; Lazaridis, T. Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys.* **2007**, *9*, 573–581.
- (7) Mishima, O.; Stanley, H. E. The relationship between liquid, supercooled and glassy water. *Nature* **1998**, *396*, 329–335.
- (8) Chaplin, M. The Anomalous Properties of Water. <http://www.lsbu.ac.uk/water/anmlies.html> (accessed Jul 30, 2011).
- (9) Hansen, J.-P.; McDonald, I. *Theory of Simple Liquids*, 3rd ed.; Elsevier: New York, 2006.
- (10) Angell, C. A.; Bressel, R. D.; Hemmati, M.; Sarec, E. J.; Tucker, J. C. Water and its anomalies in perspective: tetrahedral liquids with and without liquid–liquid phase transitions. *Phys. Chem. Chem. Phys.* **2000**, *2*, 1559–1566.
- (11) Errington, J. R.; Debenedetti, P. G. Relationship between structural order and the anomalies of liquid water. *Nature* **2001**, *409*, 318–321.
- (12) Sharma, R.; Chakraborty, S. N.; Chakravarty, C. Entropy, diffusivity, and structural order in liquids with waterlike anomalies. *J. Chem. Phys.* **2006**, *125*, 204501.
- (13) Agarwal, M.; Sharma, R.; Chakravarty, C. Ionic melts with waterlike anomalies: Thermodynamic properties of liquid BeF₂. *J. Chem. Phys.* **2007**, *127*, 164502.
- (14) Agarwal, M.; Chakravarty, C. Waterlike Structural and Excess Entropy Anomalies in Liquid Beryllium Fluoride. *J. Phys. Chem. B* **2007**, *111*, 13294–13300.
- (15) Sharma, R.; Agarwal, M.; Chakravarty, C. Estimating the entropy of liquids from atom–atom radial distribution functions: silica, beryllium fluoride and water. *Mol. Phys.* **2008**, *106*, 1925–1938.
- (16) Mittal, J.; Errington, J. R.; Truskett, T. M. Relationship between thermodynamics and dynamics of supercooled liquids. *J. Chem. Phys.* **2006**, *125*, 076102.
- (17) Jabes, B. S.; Agarwal, M.; Chakravarty, C. Tetrahedral order, pair correlation entropy, and waterlike liquid state anomalies: Comparison of GeO₂ with BeF₂, SiO₂, and H₂O. *J. Chem. Phys.* **2010**, *132*, 234507.
- (18) Agarwal, M.; Singh, M.; Sharma, R.; Alam, M. P.; Chakravarty, C. Relationship between Structure, Entropy, and Diffusivity in Water and Water-Like Liquids. *J. Phys. Chem. B* **2010**, *114*, 651–659.
- (19) Chopra, R.; Truskett, T. M.; Errington, J. R. On the use of excess entropy scaling to describe the dynamic properties of water. *J. Phys. Chem. B* **2010**, *114*, 10558–10566.
- (20) Hujo, W.; Jabes, B. S.; Rana, V. K.; Chakravarty, C.; Molinero, V. The Rise and Fall of Anomalies in Tetrahedral Liquids. 2011, arXiv:1107.5623v1 [cond-mat.soft]. <http://arxiv.org/abs/1107.5623> (accessed Jul 29, 2011).

- (21) Molinero, V.; Sastry, S.; Angell, C. A. Tuning of Tetrahedrality in a Silicon Potential Yields a Series of Monatomic (Metal-like) Glass Formers of Very High Fragility. *Phys. Rev. Lett.* **2007**, *97*, 075701.
- (22) Angell, C. A.; Kanno, H. Density Maxima in High-Pressure Supercooled Water and Liquid Silicon Dioxide. *Science* **1976**, *193*, 1121–1122.
- (23) Sciortino, F.; Poole, P. H.; Essmann, U.; Stanley, H. E. Line of compressibility maxima in the phase diagram of supercooled water. *Phys. Rev. E* **1997**, *55*, 727–737.
- (24) Scala, A.; Starr, F.; La Nave, E.; Stanley, H.; Sciortino, F. Free energy surface of supercooled water. *Phys. Rev. E* **2000**, *62*, 8016–8020.
- (25) Abascal, J. L. F.; Vega, C. Widom line and the liquid-liquid critical point for the TIP4P/2005 water model. *J. Chem. Phys.* **2010**, *133*, 234502.
- (26) Paschek, D. Temperature dependence of the hydrophobic hydration and interaction of simple solutes: An examination of five popular water models. *J. Chem. Phys.* **2004**, *120*, 6674–6690.
- (27) Lynden-Bell, R.; Giovambattista, N.; Debenedetti, P.; Head-Gordon, T.; Rossky, P. Hydrogen bond strength and network structure effects on hydration of non-polar molecules. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2748–2757.
- (28) Hummer, G.; Garde, S.; Garca, A. E.; Pohorille, A.; Pratt, L. R. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci.* **1996**, *93*, 8951–8955.
- (29) Sarupria, S.; Garde, S. Quantifying Water Density Fluctuations and Compressibility of Hydration Shells of Hydrophobic Solutes and Proteins. *Phys. Rev. Lett.* **2009**, *103*, 037803.
- (30) Acharya, H.; Vembanur, S.; Jamadagni, S. N.; Garde, S. Mapping hydrophobicity at the nanoscale: Applications to heterogeneous surfaces and proteins. *Faraday Discuss.* **2010**, *146*, 353–365.
- (31) Hess, B.; van der Vegt, N. F. A. Hydration thermodynamic properties of amino acid analogues: A systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.
- (32) Tilton, R. F., Jr.; Dewan, J. C.; Petsko, G. A. Effects of Temperature on Protein Structure and Dynamics: X-ray Crystallographic Studies of the Protein Ribonuclease-A at Nine Different Temperatures from 98 to 320 K. *Biochemistry* **1992**, *31*, 2469–2481.
- (33) Zhou, Y.; Vitkup, D.; Karplus, M. Native Proteins are Surface-molten Solids: Application of the Lindemann Criterion for the Solid versus Liquid State. *J. Mol. Biol.* **1999**, *285*, 1371–1375.
- (34) Vitkup, D.; Ringe, D.; Petsko, G. A.; Karplus, M. Solvent mobility and the protein 'glass' transition. *Nature* **2000**, *7*, 34–38.
- (35) Kumar, P.; Yan, Z.; Xu, L.; Mazza, M. G.; Buldyrev, S. V.; Chen, S.; Sastry, S.; Stanley, H. E. Glass Transition in Biomolecules and the Liquid-Liquid Critical Point of Water. *Phys. Rev. Lett.* **2006**, *97*, 177802.
- (36) Ringe, D.; Petsko, G. The glass transition in protein dynamics: what it is, why it occurs, and how to exploit it. *Biophys. Chem.* **2003**, *105*, 667–680.
- (37) Leach, A. R. *Molecular Modelling: Principles and Applications*; Addison Wesley Longman Limited: China, 1998.
- (38) Xantheas, S. S. Cooperativity and hydrogen bonding network in water clusters. *Chem. Phys.* **2000**, *258*, 225–231.
- (39) Goldman, N.; Leforestier, C.; Saykally, R. J. A 'first principles' potential energy surface for liquid water from VRT spectroscopy. *Philos. Trans. R. Soc., London A* **2005**, *363*, 493–508.
- (40) Vega, C.; Abascal, J. L. F.; Conde, M.; Aragoes, J. What ice can teach us about water interactions: a critical comparison of the performance of different water models. *Faraday Discuss.* **2009**, *141*, 251–276.
- (41) Vega, C.; Abascal, J. L. F. Relation between the melting temperature and the temperature of maximum density for the most common models of water. *J. Chem. Phys.* **2005**, *123*, 144504.
- (42) Agarwal, M.; Alam, M. P.; Chakravarty, C. Thermodynamic, Diffusional and Structural Anomalies in Rigid-body Water Models. *J. Phys. Chem. B* **2011**, *115*, 6935–6945.
- (43) Scala, A.; Starr, F.; Nave, E.; Sciortino, F.; Stanley, H. Configurational entropy and diffusivity of supercooled water. *Nature* **2000**, *406*, 166–169.
- (44) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (45) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996**, *105*, 1902–1921.
- (46) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.
- (47) Pi, H. L.; Aragoes, J. L.; Vega, C.; Noya, E. G.; Abascal, J. L. F.; Gonzalez, M. A.; McBride, C. Anomalies in water as obtained from computer simulations of the TIP4P/2005 model: density maxima, and density, isothermal compressibility and heat capacity minima. *Mol. Phys.* **2009**, *107*, 365–374.
- (48) Sanz, E.; Vega, C.; Abascal, J. L. F.; MacDowell, L. G. Phase diagram of water from computer simulation. *Phys. Rev. Lett.* **2004**, *92*, 255701.
- (49) Glass, D. C.; Krishnan, M.; Nutt, D. R.; Smith, J. C. Temperature Dependence of Protein Dynamics Simulated with Three Different Water Models. *J. Chem. Theory Comput.* **2010**, *6*, 1390–1400.
- (50) Nutt, D. R.; Smith, J. C. Molecular Dynamics Simulations of Proteins: Can the Explicit Water Model Be Varied? *J. Chem. Theory Comput.* **2007**, *3*, 1550–1560.
- (51) Agarwal, M.; Kushwaha, H. R.; Chakravarty, C. Local Order, Energy and Mobility of Water Molecules in the Hydration Shell of Small Peptides. *J. Phys. Chem. B* **2010**, *114*, 651–659.
- (52) Chau, P.-L.; Hardwick, A. J. A new order parameter for tetrahedral configurations. *Mol. Phys.* **1998**, *93*, 511–518.
- (53) Mudi, A.; Ramaswamy, R.; Chakravarty, C. Signatures of multiple time-scale behaviour in the power spectra of water. *Chem. Phys. Lett.* **2003**, *376*, 683–689.
- (54) Mudi, A.; Chakravarty, C. Effect of ionic solutes on the hydrogen bond network dynamics of water: power spectral analysis of aqueous NaCl solutions. *J. Phys. Chem. B* **2006**, *110*, 8422–8431.
- (55) Sasai, M.; Ohmine, I.; Ramaswamy, R. Long time fluctuation of liquid water - 1/f spectrum of energy fluctuation in hydrogen bond network rearrangement dynamics. *J. Chem. Phys.* **1992**, *96*, 3045–3053.
- (56) Mudi, A.; Chakravarty, C. Multiple Time-Scale Behavior of the Hydrogen-Bond Network in Water. *J. Phys. Chem. B* **2004**, *108*, 19607–19613.
- (57) Mudi, A.; Chakravarty, C. Multiple Time-Scale Behavior of the Hydrogen-Bond Network in Water. *J. Phys. Chem. B* **2006**, *110*, 4502.
- (58) Handgraaf, J.-W.; Zerbetto, F. Molecular Dynamics Study of Onset of Water Gelation around the Collagen Triple Helix. *Proteins* **2006**, *64*, 711–718.
- (59) Sinha, S. K.; Bandyopadhyay, S. Differential Flexibility of the Secondary Structures of Lysozyme and the Structure and Ordering of Surrounding Water Molecules. *J. Chem. Phys.* **2011**, *134*, 115101.
- (60) Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298K. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.
- (61) Abascal, J. L. F.; Vega, C. Dipole-Quadrupole Force ratios determine the ability of potential models to describe the phase diagram of water. *Phys. Rev. Lett.* **2007**, *98*, 237801–237804.
- (62) Rick, S. W. A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums. *J. Chem. Phys.* **2004**, *120*, 6085–6093.
- (63) Smith, W.; Yong, C. W.; Rodger, P. M. DLPOLY: Application to molecular simulation. *Mol. Simulat.* **2002**, *28*, 385–471.
- (64) Smith, W.; Forester, T. R.; Todorov, I. T. *The DL_POLY 2 User Manual 2008 Version 2.19*; STFC Daresbury Laboratory: Daresbury, U.K., 2008.
- (65) Smith, W.; Yong, C. W.; Rodger, P. M. *DL_POLY 2.19*; CSE Department, STFC Daresbury Laboratory: Daresbury, U.K., 2002.
- (66) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

- (67) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *NAMD 2.7*; University of Illinois: Urbana-Champaign, IL, 2005.
- (68) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910–8922.
- (69) Bolhuis, P. G. Transition-path sampling of β -hairpin folding. *Proc. Natl. Acad. Sci.* **2003**, *100*, 12129–12134.
- (70) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (71) Humphrey, W.; Dalke, A.; Schulten, K. *VMD: Visual Molecular Dynamics 1.8.7*; University of Illinois: Urbana-Champaign, IL, 1996.
- (72) Jhon, Y. I.; No, K. T.; Jhon, M. S. Common features of orientational order at the temperature of maximum density for various water models: molecular dynamics study. *J. Phys. Chem. B* **2007**, *111*, 9897–9899.
- (73) Jhon, Y. I.; No, K. T.; Jhon, M. S. The molecular-level relationship between the properties of liquid water molecules and orientational order. *Fluid Phase Equilib.* **2006**, *244*, 160–166.
- (74) Sharma, R.; Mudi, A.; Chakravarty, C. Diffusional anomaly and network dynamics in liquid silica. *J. Chem. Phys.* **2006**, *125*, 044705.
- (75) Sharma, R.; Mudi, A.; Chakravarty, C. Diffusional anomaly and network dynamics in liquid silica. *J. Chem. Phys.* **2009**, *130*, 199903.
- (76) Jana, B.; Bagchi, B. Intermittent Dynamics, Stochastic Resonance and Dynamical Heterogeneity in Supercooled Liquid Water. *J. Phys. Chem. B* **2009**, *113*, 2221–2224.
- (77) Bosio, L.; Chen, S.-H.; Teixeira, J. Isochoric temperature differential of the X-ray structure factor and structural rearrangements in low-temperature heavy water. *Phys. Rev. A* **1983**, *27*, 1468–1475.
- (78) Soper, A. K.; Philips, M. G. A new determination of the structure of water at 25°C. *Chem. Phys.* **1986**, *107*, 47–60.
- (79) Kumar, P.; Buldyrev, S. V.; Stanley, H. E. A tetrahedral entropy for water. *Proc. Natl. Acad. Sci.* **2009**, *106*, 22130–22134.
- (80) Bandyopadhyay, S.; Chakraborty, S.; Bagchi, B. Coupling between hydration layer dynamics and unfolding kinetics of HP-36. *J. Chem. Phys.* **2006**, *125*, 084912.
- (81) Lee, S. L.; Debenedetti, P. G.; Errington, J. R. A computational study of hydration, solution structure, and dynamics in dilute carbohydrate solutions. *J. Chem. Phys.* **2005**, *122*, 204511.
- (82) Mudi, A.; Chakravarty, C.; Ramaswamy, R. Spectral signatures of the diffusional anomaly in water. *J. Chem. Phys.* **2005**, *122*, 104507. **2006**, *124*, 069902.
- (83) Sedlmeier, F.; Horinek, D.; Netz, R. R. Spatial Correlations of Density and Structural Fluctuations in Liquid Water: A Comparative Simulation Study. *J. Am. Chem. Soc.* **2011**, *133*, 1391–1398.
- (84) Molinero, V.; Moore, E. B. Water Modeled as an Intermediate Element between Carbon and Silicon. *J. Phys. Chem. B* **2009**, *113*, 4008–4016.
- (85) Te, J. A.; Tan, M.-L.; Ichiye, T. Solvation of glucose, trehalose, and sucrose by the soft-sticky dipole-quadrupole-octupole water model. *Chem. Phys. Lett.* **2010**, *491*, 218–223.
- (86) Chaimovich, A.; Shell, M. S. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1901–1915.
- (87) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*; Academic Press: Waltham, MA, 2002.
- (88) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford University Press: New York, 1992.

Insights into Ligand–Protein Binding from Local Mechanical Response

Jagdish Suresh Patel,[†] Davide Branduardi,^{*,†} Matteo Masetti,^{†,‡} Walter Rocchia,[†] and Andrea Cavalli^{*,†,‡}

[†]Department of Drug Discovery and Development, Italian Institute of Technology, Via Morego 30, I-16163 Genova, Italy

[‡]Department of Pharmaceutical Sciences, Alma Mater Studiorum, Bologna University, Via Belmeloro 6, I-40126 Bologna, Italy

 Supporting Information

ABSTRACT: Computational studies of ligand–protein binding are crucial for properly designing novel compounds of potential pharmacological interest. In this respect, researchers are increasingly interested in steered molecular dynamics for ligand–protein binding and unbinding studies. In particular, it has been suggested that analyzing the work profiles along the ligand–protein undocking paths could be fruitful. Here, we propose that small portions of work profiles, termed “local mechanical responses” of the system to a steering force, could serve as a universal measure for capturing relevant information about the system under investigation. Specifically, we first collected a high number of steering trajectories using two biological systems of increasing complexity (i.e., alanine dipeptide and (*R*)-roscovitine/CDK5 complex). Then, we devised a novel postprocessing tool to be applied to the local mechanical responses, to extract structural information related to the biological processes under investigation. Despite the out-of-equilibrium character of the trajectories, the analysis carried out on the work profiles provided pivotal information about the investigated biological processes. This could eventually be applied to drug design.

INTRODUCTION

Computational drug design has two major goals: (i) the accurate estimation of ligand–protein binding free energy; and (ii) the disclosure of the structural determinants responsible for ligand–protein recognition and binding. In this scenario, molecular dynamics (MD)-based enhanced sampling methods play an increasingly relevant role. As far as the ligand–protein binding free energy (ΔG_b) is concerned, the most widely used strategies are based on alchemical transformations¹ (double decoupling^{2,3} and related schemes), where the ΔG_b is estimated using a thermodynamic cycle.^{1,4–7} Notwithstanding the impressive results achieved, alchemical transformations^{8,9} do not explicitly account for the dynamical events occurring upon ligand–protein recognition and binding. Notably, such dynamical events can be of paramount importance in drug discovery by providing fundamental drug design information.¹⁰ Moreover, simulations of the unbinding process are relevant for the binding kinetics, and thus for drug residence time within the target.^{11,12} Very recently, the first promising attempt to obtain the kinetics from straightforward simulations appeared¹³ and required a collective computational effort through distributed networks. Therefore, for the time being, umbrella sampling,^{14,15} metadynamics,^{16,17} and steered MD¹⁸ are the methods of choice to disclose the structural determinants relevant to a ligand binding to a protein on the exit pathways. In particular, metadynamics has proven to be rather effective,^{17,19} but its computational cost is not a priori predictable even in an exploratory regime (i.e., when convergence of the free energy is not required). Steered MD is also becoming very popular for studying biophysical processes.^{20–26} This is partly due to its conceptual simplicity and integration in several currently available MD codes. In steered MD, a certain transition (such as ligand–protein unbinding) is obtained via a tunable

restraining potential, which forces the system to move away from its initial configuration (e.g., a bound state for ligand–protein complexes) to a given position during an MD run.¹⁸ In the case of an unbinding process, the target position may be defined as an unbound state where the ligand–protein interactions may be considered negligible. In steered MD, the simulation time required to complete the transition is an input parameter, which can be reduced to an almost arbitrary small number, making the technique particularly appealing in the drug discovery process. Moreover, some recent attempts to obtain ΔG_b of binding have been reported,²⁵ but these are limited to very simple model ligands. Concerning more realistic cases, Colizzi et al.²⁴ have demonstrated that steered MD can be applied to drug design-related problems without requiring an accurate estimation of the ΔG_b . In particular, Colizzi et al. could discern active from inactive enzyme inhibitors by a simple visual inspection of the force profiles required for pulling ligands out of the protein binding site. Increasing the number of steered MD pulling trajectories provides a natural extension to improve the overall reliability of this approach,^{26–28} and this is becoming accessible due to the ongoing increase in CPU performance. Although much recent effort has been devoted to analyzing configurations generated by MD runs,^{29–31} it is not easy to extract relevant structural information from an ensemble of steered MD work profiles for a complex system. This is due to the lack of specific and effective analysis tools.

Here, we report on a novel postprocessing strategy aimed at analyzing steered MD trajectories and extracting the structural features that are relevant to the biological process under investigation.

Received: May 12, 2011

Published: August 30, 2011

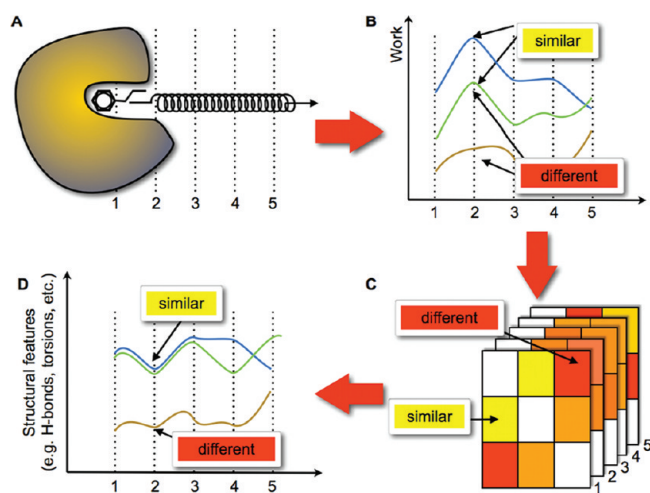


Figure 1. A sketch of the general workflow used in this study. Steered MD simulations are carried out (A), work profiles are subsequently generated (B), and a set of matrices that record the similarity among profiles are produced (C). The matrices are then postprocessed, and the similarities in the work curves are correlated with structural elements (D).

We used the work profile calculated on the system in a limited region of space, hereafter referred to as local mechanical response (LMR), as a measure of the ligand–protein interaction strength. The LMR profiles were analyzed via multidimensional scaling (MDS),³² which allowed us to extract structural similarity/dissimilarity over a large set of steered MD-derived LMRs. We also report on using this postprocessing tool to correlate work profiles with the relevant structural changes occurring during a certain process (see Figure 1 for a schematic representation of the overall process). In particular, we show that: (i) reasonably fast pulling regimes can provide relevant information on structurally different reactive pathways; (ii) by analyzing LMR patterns, it is possible to point out the structural elements important for the unbinding process; (iii) LMR emerges as a simple and system-independent observable that has general applicability for ligand–protein binding studies. To illustrate the usefulness of this novel approach, two case studies of increasing complexity are here investigated.

METHODS

Work Estimate in Steered MD Simulations. In the thermodynamic integration formalism,³³ the free energy difference associated with the continuous change of the system from an initial state described with an Hamiltonian $H(\lambda_0)$ to a final one $H(\lambda_1)$ obtained by changing the parameter λ is given by

$$\Delta G = \int_{\lambda_0}^{\lambda_1} \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle d\lambda \quad (1)$$

One possible way to achieve the transition is to add to the standard Hamiltonian a harmonic potential $U_{\text{bias}}(t, \mathbf{r})$ acting on a descriptor $s(\mathbf{r})$ (e.g., the ligand–protein distance or the mean square deviation with respect to a given structure), which holds the following time dependency:

$$U(t, \mathbf{r}) = k(s(\mathbf{r}) - s_0 - vt)^2 \quad (2)$$

where s_0 is the value of the descriptor in the initial state $H(\lambda_0)$, t is the time, and k is a numerical constant. Thus, whenever time is considered in place of the parameter λ , the partial derivative of the Hamiltonian turns out to be the instantaneous value, and the integral in eq 1 (which can be easily calculated via quadrature or trapezoidal rule) corresponds to the work ΔW exerted on the system:

$$\Delta W = \int_{t_0}^{t_1} \left\langle \frac{\partial H(t)}{\partial t} \right\rangle dt = -2kv \int_{t_0}^{t_1} (s(\mathbf{r}) - s_0 - vt) dt \quad (3)$$

After a predetermined amount of time, the center of the harmonic constraint will be located in its final position:

$$s_1 = s_0 + vt_1 \quad (4)$$

Therefore, whenever the spring constant k chosen is large enough (stiff-spring regime), it is reasonable to assume that, at the final time t_1 , the system has approximately reached the point s_1 . Moreover, when the $U_{\text{bias}}(t, \mathbf{r})$ is applied in a quasistatic regime (slow growth), the calculated work is equivalent to the free energy estimate obtained by thermodynamic integration:²⁸

$$\Delta W(t_1, t_0) \cong \Delta G(s_1, s_0) \quad (5)$$

This is a translation of the classical result, which states that the work exerted by an external potential to move a system quasistatically from an initial to a final state is equal to the free energy difference between these two states.

From a practical standpoint, Crooks theorem^{27,34} or Jarzynski²⁷ equality are better suited to evaluate free energies from out-of-equilibrium trajectories, since the quasistatic limit is practically never reached. Thus, in such cases the work profile turns out to be rather different from the actual free energy landscape. Here, we explore the hypothesis that, even in a non-quasistatic regime, the work profiles retain some information about the structural events associated with the mechanical response induced by the steering procedure. Hereafter, this will be referred to as the “mechanical response” profile.

To better understand the effects of the application of a steering potential, it is useful to consider the limiting case where the transition is carried out in a single step. During such a time span (typically 2 fs for a classical MD simulation), the system can be considered frozen, and therefore the estimated work will be

$$\Delta W = k(s_1 - s_0)^2 \quad (6)$$

In this case, the path followed by the system (e.g., ligand and protein in the case of an undocking experiment) becomes irrelevant, and all of the contribution to the work comes from the restraint position. Moreover, being system independent, the estimated work cannot be by any means similar to the free energy of binding, which, in contrast, is strongly system dependent. This conceptual experiment shows that, although the requirement of quasistatic transformation can be somewhat relaxed, the simulations should nevertheless be performed with a steering velocity that can capture the mechanical response of the system subjected to the typical relaxation time of the variables implied in the investigated process. In accordance with this limit, the ligand acts as a probe for ligand–protein interactions, and the trajectories may retain some physicochemical relevance. This pulling velocity range is worth exploiting. This is because it is computationally affordable and preferable to both very fast steering regimes (which are not sensitive to the ligand–protein rearrangement) and very low velocities (which are required whenever accurate

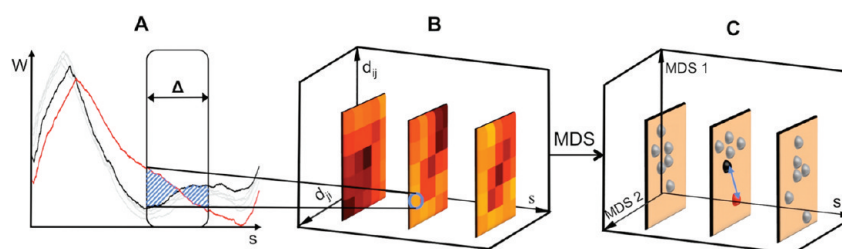


Figure 2. Schematic representation of the postprocessing analysis of the work curves. The first step (A) is the calculation of the distance between mechanical responses recorded on a small portion of the order parameter s (blue-shaded region). The second step (B) is the collection of the distance matrices along the order parameter, and the last step (C) is the retrieval of a fictitious representation via MDS in a reduced dimensionality for each of the matrices. MDS1 and MDS2 are the fictitious coordinates produced by the MDS algorithm.

free energy estimations are needed). The latter case in particular must meet a stricter requirement: Whenever multiple paths are competing, the correct free energy is obtained only when the relative occurrence of different pathways is at convergence. This is particularly unlikely when each path is separated from the others by a large barrier. In the present study, we show that structurally relevant information about the different paths can be gathered even without obtaining full convergence on the statistics among competing pathways. This results in a large computational saving.

MDS Analysis. Since a detailed analysis of a considerable number of trajectories may be cumbersome, we want to test the hypothesis that adopting a postprocessing technique for the mechanical response profiles $W(s)$ may be helpful in capturing the principal structural trends in the collected ensemble of trajectories. The rationale is that structural features are expected to affect the resistance that the system opposes to the external force and, in turn, the mechanical response profiles $W(s)$. Although reasonable, this assumption must be kept in mind while postprocessing two similar work profiles, since this does not exclude the possibility that they represent two different pathways.

In the stiff spring regime we can assume that $s(t_1) \approx s_1$ and can define the n -th mechanical response profile^{18,28} as

$$W_n(s_1) \cong W_n(t_1) = -2kv \int_{t_0}^{t_1} (s(\mathbf{r}) - s_0 - vt) dt \quad (7)$$

A portion of the mechanical response within a certain interval of s , of width Δ , is termed here “local mechanical response” (LMR). The distance between two LMRs obtained from two steered MD trajectories within a certain interval of s , of width Δ , may be defined as

$$d_{nm}(s_0, \Delta) = \int_{s_0}^{s_0 + \Delta} [W_n(s) - \langle W_n(s) \rangle_{\Delta} - W_m(s) + \langle W_m(s) \rangle_{\Delta}]^2 ds \quad (8)$$

with $\langle W_n(s) \rangle_{\Delta}$ being the average work over the interval:

$$\langle W_n(s) \rangle_{\Delta} = \frac{\int_{s_0}^{s_0 + \Delta} W_n(s) ds}{\Delta} \quad (9)$$

This defines a set of distance matrices along the pulling coordinate that can be exploited to monitor different families of LMRs along the steering pathway. A brief sketch of this step is

represented in Figure 2A and B. One valuable outcome of calculating the local difference between two profiles over blocks of fixed size Δ (instead of calculating it on the work profile over the whole transition) is the crucially important possibility of detecting branching in the mechanical response intersection among different realizations. To avoid irregular behaviors at the borders between adjacent blocks, some degree of overlap between them is allowed. Clearly, the window size Δ must be properly tuned in accordance with the scales of the events involved in the transition. For the ligand unbinding event investigated here, Δ was optimally sized to identify hydrogen (H)-bonding breaking/formation or local conformational rearrangements, while much finer thermal motions were chosen to be averaged out.

For the sake of completeness, we note that the same information could be obtained by averaging the exerted forces rather than using work. Here, we prefer to consider the work profiles because they may provide an interesting view of the dissipative work produced and a useful hint concerning the appropriateness of the pulling parameter as well as the amount of orthogonal degrees of freedom interfering with the pulling direction.

Once the set of matrices $d_{nm}(s, \Delta)$ is obtained, there are two possible ways to postprocess them. One approach relies on standard clustering techniques. Because of the intrinsic need for a clustering threshold value whose choice in the case of the work curve realizations would be arbitrary and far from simple, we decided to adopt a MDS approach^{29,32} so as to have a direct grasp over the topology of the LMR pattern. Hence, we term this analysis LMR-MDS.

MDS is a standard pattern-recognition technique³² that can detect the intrinsic dimensionality from a distance matrix by searching for the lowest dimensional possible manifold that can preserve the reciprocal neighbor distances and produce a representation that resembles the original topology in the original high-dimensional space.

The simple MDS algorithm used here initially picks a random entry $d_{nm}(s, \Delta)$ (i.e., the distance between the n and m work profiles over an interval of width Δ centered in a given point s along the steering coordinate) in the high M -dimensional original distance matrix $d(s, \Delta)$ and creates two fictitious points in a low D -dimensional Euclidean space at the same distance $d_{nm}^{\text{MDS}}(s, \Delta)$. Then, another entry on the same row n but at a different column l is picked ($d_{nl}(s, \Delta)$) and used to constrain the distance from n when a new point representing l is placed in the D -dimensional space. Since this operation is not univocal, a Monte Carlo (MC) procedure is used to satisfy as much as possible all the distances with respect to the already projected elements. The procedure is iterated until the entire original distance matrix $d(s, \Delta)$ is spanned. At the end, the positioning of

the last points is subject to a larger number of constraints than the initial ones. To reduce the residual strain, we therefore applied a MC annealing procedure, adopting the merit function $E_{MC}(s, \Delta)$, also called “stress” function in MDS terminology, to all the points in the D-dimensional representation. The stress function $E_{MC}(s, \Delta)$ is the squared difference in the representation of the two matrices, the M-dimensional and its fictitious D-dimensional representation:

$$E_{MC}(s, \Delta) = \sum_{n < m}^N [d_{nm}(s, \Delta) - d_{nm}^{MDS}(s, \Delta)]^2 \quad (10)$$

In order to evaluate the appropriateness of the choice of D, the preservation of an order relation within the distance matrices is verified. In the negative case, the dimension is increased, and the entire procedure is repeated.

The postprocessing procedure consists of applying the above procedure for each interval Δ along the steering coordinate. The final outcome is thus a set of reduced representations for this degree of freedom (see Figure 2C).

At the end of the procedure, one gets an intuitive representation of the different families of steering processes that may occur. By inspecting the structural differences between representative members of different families, we obtain a picture of the different processes occurring at the molecular level during ligand unbinding. This final step is represented in Figure 1C and D.

We note that a similar postprocessing tool can be applied to any set of distance matrices. Therefore, when studying the alanine dipeptide, we also applied MDS to the Cartesian coordinates of the atoms. In this case, each entry of the matrix $d_{nm}^{RMSD}(s, \Delta)$ was obtained using the root mean square deviation (RMSD) of the heavy atoms of two structures along the steering path, obtained after optimal alignment through the Kearsley algorithm.³⁵ We term this analysis “Cartesian-MDS” to differentiate it from the work-based LMR-MDS. Their comparison (see Results Section) is instrumental in verifying the connection between the local mechanical response and the structural changes of the system.

It is worth noting that the MDS implementation here described was intentionally rather unsophisticated, and for this specific problem that included a data set with relatively modest size, our MDS approach could be basically equivalent to other more advanced methods, like classical Torgerson multidimensional scaling. The latter approach has to be highly recommended when a larger data set has to be analyzed.³²

Simulation Details. All the MD simulations in the present work were carried out with NAMD2.7 code.³⁶ The simulations in the NVT ensemble were performed using the Langevin thermostat,³⁷ and additional steering forces were introduced via the PLUMED³⁸ plugin integrated in the NAMD code.

For alanine dipeptide (see Figure 3A for a molecular sketch), we used CHARMM27 force field,³⁹ a time step of 0.2 fs without constraining the covalent-bond length involving hydrogen atoms so as to maximize the number of degrees of freedom involved. A Langevin thermostat at 300 K with relaxation time of 8 ps was used to maintain the average temperature during out-of-equilibrium runs. We performed steered MD simulation using mean square displacement (MSD) of the heavy atoms with respect to C_{7ax} configuration as a pulling coordinate. Optimal alignment was obtained using the Kearsley algorithm.³⁵ A number of tests were performed to choose the speed for pulling and the magnitude of the spring constant. A limited dissipative work was obtained with a value of 2000 kcal/($\text{\AA}^4 \cdot \text{mol}$) for the spring constant and 0.005 $\text{\AA}^2/\text{ps}$

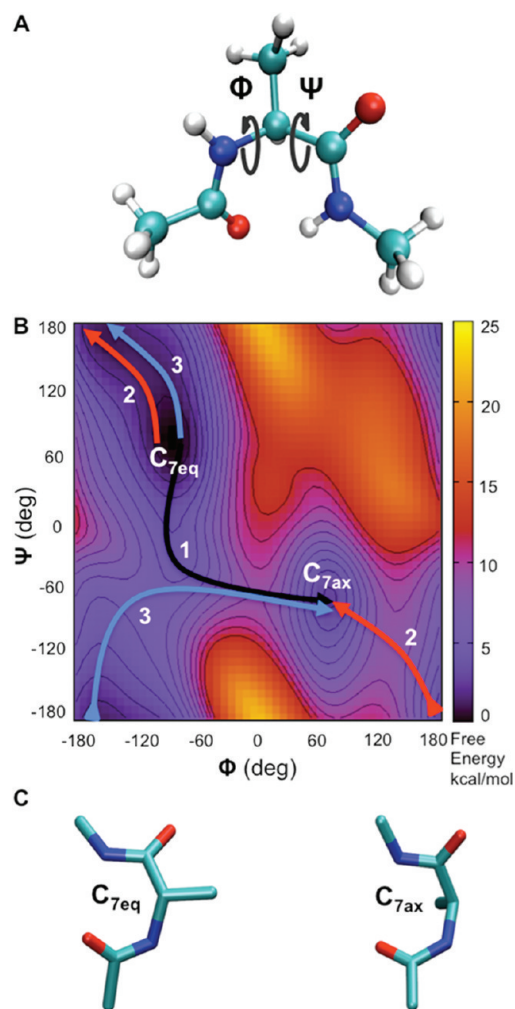


Figure 3. Structural and free energy features of alanine dipeptide. (A) Ball and stick representation of alanine dipeptide along with the Φ and Ψ dihedrals used in the Ramachandran plot are shown. (B) Free energy landscape of alanine dipeptide as a function of the Φ and Ψ dihedrals produced via umbrella sampling. The isoline separation is of 1.0 kcal/mol. Both C_{7eq} and C_{7ax} minima are connected by three different pathways denoted with three different colors (black, red, and blue). (C) Stick representation of the two metastable conformers C_{7eq} and C_{7ax} (heavy atoms only).

for the pulling speed (Supporting Information, Figure S1). Unless specified, all the molecular representations were generated using VMD.⁴⁰

For (*R*)-roscovitine/CDK5 complex, the starting geometry used in the simulations was obtained after removing the p25 activator from the X-ray structure retrieved from the Protein Data Bank (PDB code: 1UNL).⁴¹ The Amber parm99SB⁴² force field was used for the protein, while the (*R*)-roscovitine was treated with the general Amber force field for organic molecules⁴³ and the charges were derived according to the restrained electrostatic potential (RESP) procedure.⁴⁴ Prior to the steered MD simulations, the system was minimized and equilibrated in a box with 10 371 TIP3P⁴⁵ water molecules and pressurized for 2 ns in the NPT ensemble using a Langevin thermostat³⁷ and a Langevin piston barostat.⁴⁶ Long-range electrostatic interactions were treated with the particle mesh Ewald (PME) method.⁴⁷ Short-range nonbonded interactions were calculated using a cutoff

radius of 12 Å for both Coulomb and van der Waals potentials. A 2 fs time step was used, and covalent-bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm.⁴⁸

Undocking experiments were performed by means of steered MD. The pulling variable was the distance between the center of mass of the residues belonging to the binding site of the protein and the center of mass of the ligand (see Supporting Information, Figure S2 for a pictorial view of the binding pocket). The pulling parameters (spring constant, pulling velocity, and maximum extension) were determined by performing different trial simulations in different conditions and by comparing the calculated work values. In particular, 12 steered MD runs were performed using a spring constant of 10, 100, 1000 kcal/(mol·Å²) and pulling velocities of 0.5, 0.2, 0.1, and 0.01 Å/ps. In the end, a spring constant of 100 kcal/(mol·Å²) along with a pulling velocity of 0.01 Å/ps were chosen, since they provided the lowest work values in stiff spring regime. The target distance for the steered MD simulations was chosen to be 22 Å. This is because preliminary runs showed that, at this distance, the ligand was completely detached from the target (Supporting Information, Figure S3). Moreover, this pulling velocity was shown not to irreversibly disrupt the secondary structure (Supporting Information, Figure S4).

RESULTS AND DISCUSSION

The content of this section is here briefly outlined. First, the postprocessing technique was tested on a set of steered MD runs using alanine dipeptide in vacuum. This is a widely used test case for benchmarking various MD-based simulation techniques. Then, the procedure was applied to the pharmacologically relevant complex (R)-roscovitine/CDK5, which is currently being investigated in the search for novel drug candidates to treat Alzheimer's disease.

Alanine Dipeptide. Our prototypical case study is the alanine dipeptide in vacuum (see Figure 3A), a system widely used as a benchmark for enhanced sampling schemes.^{15,16,49–52} Alanine dipeptide, in a lower dimension, recapitulates most of the features that are relevant to the free energy landscape of ligand–protein recognition and binding. In particular, it displays two minima: a wider basin that can resemble a ligand in the bulk of the solvent, and a much narrower basin that can be similar to a ligand–protein complex. These two minima are connected by multiple reactive pathways with sizable free energy barriers (see Figure 3B). The two main metastable states are usually denoted as C_{7eq} and C_{7ax} (see Figure 3C), and the transition from one to the other is generally represented in terms of the dihedral angles phi (Φ) and psi (Ψ) (Figure 3A and B). Due to the periodic nature of these descriptors, three distinct pathways connect C_{7eq} to C_{7ax} and two different free energy barriers can be identified (8.5 and 10.5 kcal/mol; see Figure 3B). Furthermore, alanine dipeptide represents an ideal test case since it allows steered MD to be performed in a fully reversible work regime. This situation is highly desirable as it allows us to minimize the effects of energy dissipation. In contrast, protein–ligand unbinding processes display a large number of degrees of freedom that are only partly orthogonal to the pulling direction, thus showing a much larger relaxation time. As a result, one usually observes a significant increase in the amount of dissipative work, with respect to the alanine dipeptide case study, and the distribution of work values is broader.

Alanine dipeptide was initially investigated by simple MD, which allowed us to generate an ensemble of starting structures in

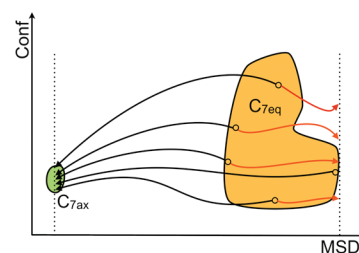


Figure 4. A sketch of the extension procedure adopted for the steered MD experiments in alanine dipeptide. The starting ensemble is represented in orange (C_{7eq}), while the targeted ensemble is in green (C_{7ax}). The starting configurations may assume different positions in the MSD variable. So as not to be limited by the shortest pulling experiment (black arrows), the pulling was extended to the highest MSD value occurring in the equilibrium ensemble (red arrows).

C_{7eq} basin to be used for subsequent steered MD runs. MSD was then used (see Methods Section) as the steering variable s to drive the system from C_{7eq} to C_{7ax} within the targeted MD framework.⁵³ About 500 simulations were carried out starting from initial configurations to provide a fairly large initial data set of trajectories for building a robust statistics for all three pathways. Since each of the possible paths from C_{7eq} to C_{7ax} may start from a different point in the C_{7eq} basin (Figure 4), the trajectories displayed variable length. To consistently compare them, the steered MD of the shorter trajectories was elongated in a backward direction to cover the same MSD range (overall 2.8 Å²) as the longest one. These additional steered MD trajectories are represented by red arrows in Figure 4. This allowed us to produce a set of mechanical response profiles spanning homogeneously from 0.0 to 2.8 Å², corresponding to the C_{7ax} and C_{7eq} conformation, respectively. The ensuing profiles covered all three possible pathways from C_{7eq} to C_{7ax} . Of these, we considered just 60 work profiles (20 for each pathway) to minimize redundancy within the set. We could thus check the ability of our postprocessing tool to detect the structural features relevant to each reactive pathway and control the reliability of the results. This was possible because structural differences along the C_{7eq} – C_{7ax} transition path are well-known and can easily be traced within the Ramachandran plot.⁵⁴

The work profiles were then used to compute distance matrices along the path (see Methods Section and Figures 1 and 2). The window size Δ was set to 0.80 Å² in MSD space. This value was chosen by visually inspecting the mechanical response profiles and selecting the MSD interval in which sizable structural motions occurred. A series of matrices ($d_{mm}(s, \Delta)$) was generated by calculating pair-wise distances between two work profiles (in the LMR space) from a maximum of 2.80 Å² to a minimum of 0.04 Å² by steps of 0.28 Å² (roughly equal to $\Delta/3$). These matrices were then processed by the MDS technique, as reported in the Methods Section. Here, a single dimension was sufficient. This is because additional dimensions did not significantly reduce the difference between the reduced and the full dimensionality distance matrices.

Figure 5A shows the projections of the steered MD trajectories over the Ramachandran plot. One representative path for each family is highlighted with different symbols and color codes. From the plot, it is evident that path 3 (blue path) started very close to path 2 (red path) in the upper left corner (around $\Phi \approx -2.0$ and $\Psi \approx 3.1$ in Figure 5A). By crossing the periodic image, it joined path 1 (black path) at $\Phi \approx -1.0$ and $\Psi \approx 0.0$, finally ending up in C_{7ax} . This sequence of events is fully

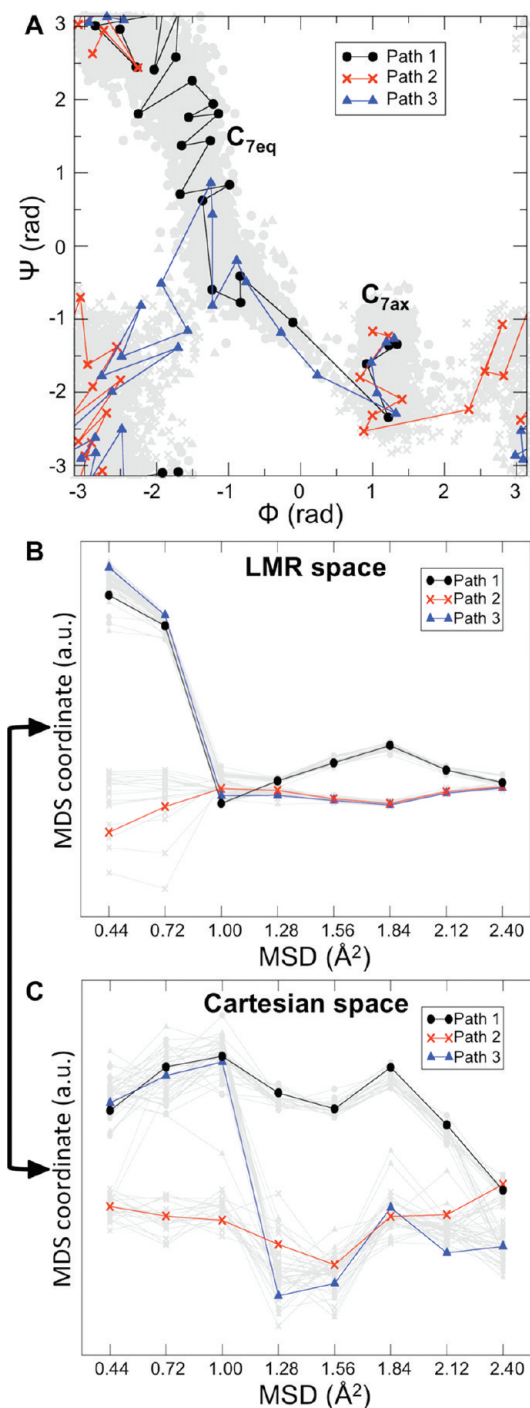


Figure 5. Analysis of alanine dipeptide trajectories. Projections of the steered MD trajectories on the Ramachandran plot (A), LMR-MDS as a function of the steering coordinate (B), and Cartesian-MDS as a function of the steering coordinate (C). Arbitrary units are used for the ordinate in the MDS representation because the coordinates from the MDS representations are fictitious. To guide the eye, in all plots, single representative trajectories for each of the three pathways are shown in black, red, and blue.

represented in the LMR-MDS of Figure 5B: path 2 and path 3 initially overlapped, and at $\text{MSD} \approx 1.00 \text{ \AA}^2$, path 3 joined path 1.

We performed the same analysis using a Cartesian-MDS. It is important to note here that the comparison between Cartesian

and LMR spaces is instrumental in verifying the connection between structural features and mechanical response. Additionally, we can test the ability of the MDS algorithm to reproduce multidimensional topologies without introducing aberrations in the context of molecular simulations.

In the Cartesian-MDS (Figure 5C), we observed a very similar pattern of events and a remarkably similar distinction in the three pathways. This strengthened our hypothesis that local mechanical response analysis can point to different structural features along several dynamical paths. Indeed, LMR-MDS enhances the distance in those points where the difference in terms of mean force is higher. In our example, two different downhill access routes to $C_{7\text{eq}}$ in the mean force space (Figure 5B) emerged as two distinct routes, even though their structures were similar at low MSD values, as reported in Figure 5C. Furthermore, small distances in LMR could recognize two paths as identical, even though they may display different structural features (around 1.0 \AA^2 in the MSD space).

In summary, although LMR-MDS does not produce an identical representation to that of Cartesian-MDS, it can convey the same information in terms of pathway topology. This implies that the raw information from the mechanical response contains relevant details of the structural rearrangement acting on the system. This could be of paramount importance for drug design. By reverting this procedure, one can postprocess the raw information contained in the work profiles so as to highlight different classes of unbinding pathways, which should point to different structural features.

In fact, in contrast with the alanine dipeptide case study, ligand–protein systems usually show a complex and dynamical network of interactions. This is because the atoms involved in the exiting pathway change along the route. As such, the structural clustering of the ligand alone may not provide enough information about the unbinding mechanism (see Supporting Information, Figure S5 for a Cartesian-MDS for ligand–protein unbinding case). Therefore, we suggest that LMR (an atom-independent observable) could be exploited to point to structural differences between trajectories for complex undocking studies.

(R)-roscovitine/CDK5 Complex. Our protocol was then applied to the cyclin-dependent kinase 5 (CDK5) in complex with the inhibitor (R)-roscovitine. CDK5 is an important kinase protein that regulates various processes in developing adult neurons. It is associated with several neural functions.^{55–58} Increased CDK5 activity has been implicated in certain neural disorders,^{59–61} including Alzheimer’s disease.⁵⁹ For these reasons, we chose the (R)-roscovitine/CDK5 complex as a pharmaceutically relevant system for the application of our postprocessing technique.

Fifty steered MD runs were performed, using as a steering coordinate the distance between the center of mass of the CDK5 binding site (the backbone heavy atoms of the residue represented in cyan in Figure 6) and the center of mass of the ligand (see Methods Section and Figure 6). The data set of 50 work profiles was then postprocessed using the LMR-MDS analysis described above. In the distance range 8–22 Å, a window size Δ of 1.5 Å and a shift factor of 0.5 Å were used to generate the matrices for MDS. As with previous calculations, a monodimensional MDS was sufficient to obtain an acceptable error in the reduced representation. In Figure 7, we show the MDS position of the points as a function of the steering coordinate. All the members in Figure 7 are color-coded according to the difference of calculated work in the considered interval Δ (see the caption of Figure 7 for further details). In particular, all the trajectories

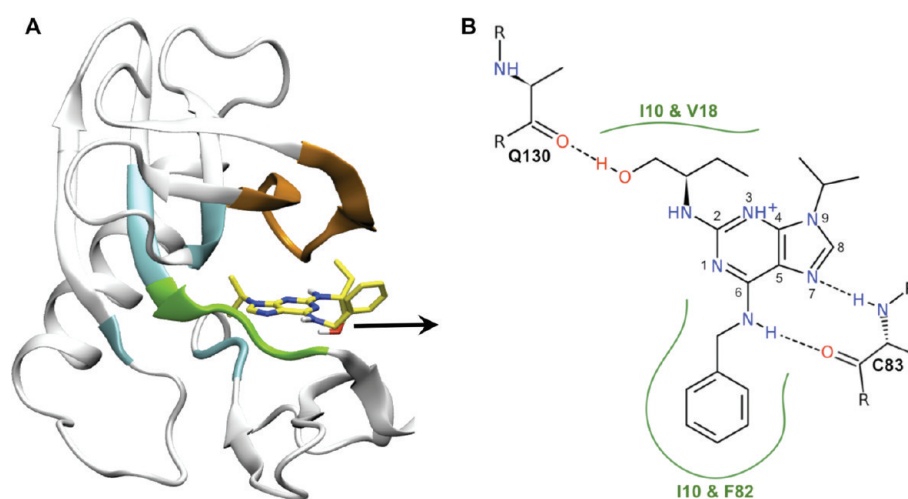


Figure 6. Three-dimensional structure of the CDKS binding pocket bound with (*R*)-roscovitine inhibitor. (A) The hinge region is highlighted in green, while the glycine-rich binding loop is highlighted in orange. (*R*)-roscovitine (stick representation) is highlighted in a C-yellow, O-red, N-blue, and H-white color scheme. The part of the binding pocket highlighted in cyan comprises all the atoms used in defining the pulling variable. The black arrow approximately indicates the steering direction. A 2D sketch of the interactions occurring between (*R*)-roscovitine and CDKS at the initial step of the steering procedure is also reported (B) and was produced using PoseView.⁶⁵

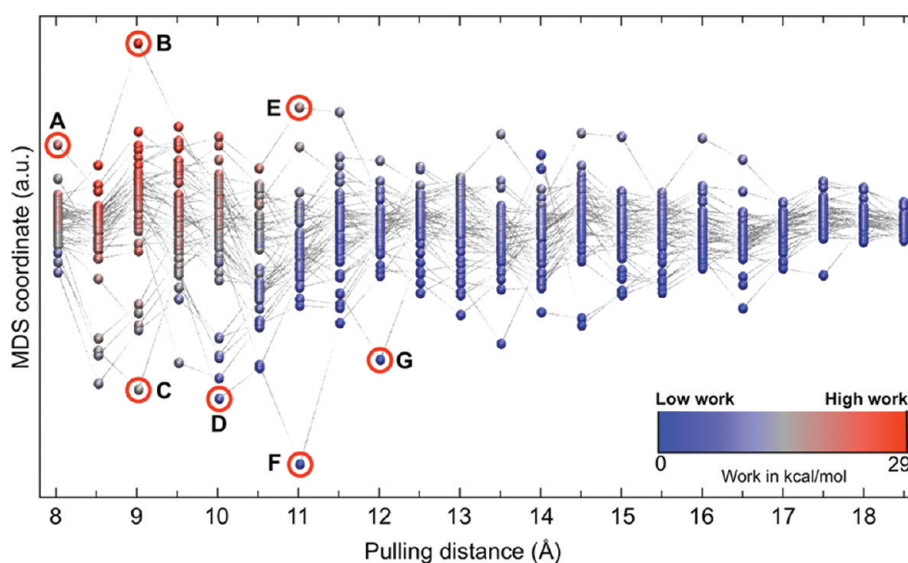


Figure 7. MDS representation of the LMR plotted against the steering coordinate obtained from the (*R*)-roscovitine/CDKS system. The work curves are shown according to a BWR color scheme, where the color of each MDS point reflects the work difference within Δ size interval. Labels refer to particular points corresponding to frames selected along the steering variable, which have been analyzed further.

seemed to follow a similar stream, and the work exerted to pull the ligand out of the pocket was primarily spent in the early stages of the unbinding process (red dots at low pulling distance in Figure 7).

In the initial structure of the complex, the NH group of Cys83 donated a H-bond to the purine nitrogen in position seven of the ligand, whereas nitrogen of benzylamino group acted as an H-bond donor to the main chain oxygen of Cys83 (Supporting Information, Figure S2 and Figures 6B and 8A). The chiral hydroxyethyl substituent of (*R*)-roscovitine interacted with Gln130 via H-bond, while the ethyl group engaged two hydrophobic interactions with Ile10 and Val18. The bulky benzyl substituent protruded into a hydrophobic pocket lined by Ile10 and Phe82 and toward the bulk of the solvent.⁴¹

During the steered MD simulations, we first observed that the interactions of the ligand with Cys83 could be lost by rotating the purine ring toward the glycine-rich loop so as to disrupt the H-bond with the protein backbone in the initial stage of unbinding. By color coding the ligand configuration according to the LMR values (see Figure 8C), we realized that this particular rotation was associated with a lower LMR, as compared to those trajectories that retained these tight interactions in the early unbinding stage. This rupture was compensated for by the formation of H-bonds between the hydroxyethylamine group of (*R*)-roscovitine and the backbone oxygens of different residues (i.e., Ile10, Glu12, Gln130, and Asp86), depending on the generated unbinding trajectories. Moreover, the oxygen of the hydroxyethyl group was found to H-bond with Lys89, Gln85, and Gln130.

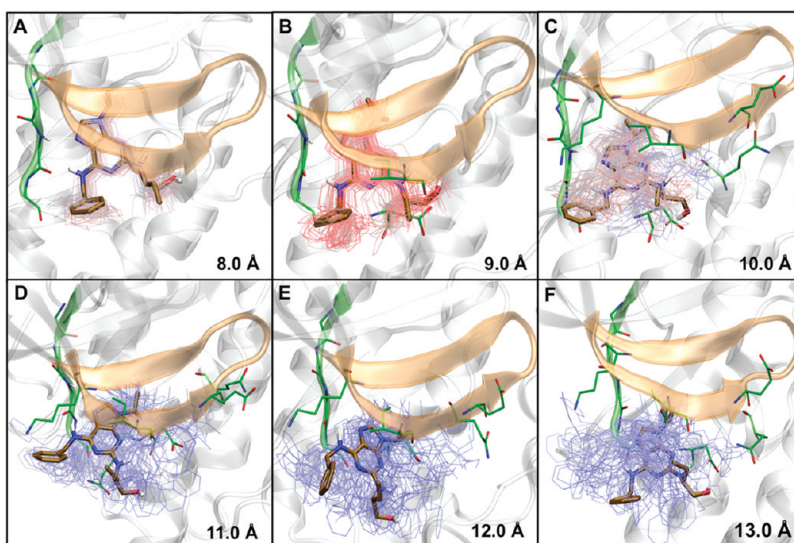


Figure 8. Representative snapshot of the exit route of the ligand from the binding pocket. The ligands represented in stick correspond to the central pose along the MDS coordinate in Figure 7. The thin lines represent the heavy atoms of the ligand, colored according the LMR measured and displayed in Figure 7.

Here too, the residues that interacted with roscovitine depended on the unbinding trajectories being investigated.

Conversely, we detected higher LMR values at the initial stage whenever the interactions with Cys83 were retained for a longer time. In this case, the ligand was found not to rotate but to favor a closer interaction with the hinge region. Here, the hydrogen donor of the benzylamine group interacted with Asp84, while charged nitrogen of Lys20 could form a cation– π interaction with the benzyl substituent of (*R*)-roscovitine. Additionally, the phenyl ring of the Phe82 was found to occasionally form T-stacking interactions with the benzyl substituent. On the other side, the H-bond interaction, between the hydroxyethyl group and the backbone oxygen of Gln130, was lost and replaced by new temporary polar interactions with the side chain of Asp86 (see Figure 8C) or backbone oxygen of Ile10 (depending on the trajectory). By coloring the ligand according to the LMRs values and aligning the configurations corresponding to a given distance, we produced an intuitive description of the energetics involved in the various interactions between the ligand and the target (Supporting Information, Figure S6 for an enhanced version of the picture).

In a later stage of the unbinding process, while the ligand approached the solvent, the amount of work decreased (blue dots at high pulling distance in Figure 7 and blue lines for conformations in Figure 8E and F) due to unspecific solvent interactions.

Then, we analyzed in depth the outlier configurations, which were labeled with red circles in Figure 7. These corresponded to values of extremely low or high LMR with respect to the central stream of the trajectories. According to our hypothesis, these outliers in the LMR space should correspond to peculiar conformations, showing specific (de)stabilizing contributions to the unbinding process that would otherwise be overlooked. As expected, most of the structurally relevant outliers were observed at small distance values, when the ligand started to move out of the protein entrance.

At a distance of 8.00 Å, the point labeled with A (see Figures 7 and 9) was characterized by a significantly well-directed H-bond pattern with Cys83. This is in agreement with the fact that

removing such a well-directed H-bond network has a large energetic cost. At a subsequent stage, the formation of an H-bond with Cys83 and the coincidental interaction of hydroxyethylamine with Asp86 were detected (point B in Figure 7 and Figure 9). This seemed plausible since the formation of this interaction induced a rotation of the ligand that could more tightly interact with the pocket, thus preventing the unbinding from taking place. Conversely, a relevant decrease of LMR seemed to be associated with a conformational rearrangement in the hinge region of the protein (see point C in Figures 7 and 9). This could be ascribed to steering induced protein deformation, which facilitated the ligand unbinding. This highlights an additional feature of our postprocessing approach, namely the possibility of detecting the mechanical response of both the protein and the ligand, thus highlighting the induced rearrangement that occurs upon ligand unbinding.

To prove this, we repeated an identical simulation by applying an RMSD harmonic restraint on those residues involved in such rearrangements on the protein backbone. In this way, the decrease of LMR was not observed anymore, and the trajectory displayed a profile very similar to the one observed in the majority of cases (Supporting Information, Figure S7).

Similarly, the point denoted with D was characterized by a remarkably low LMR with respect to the majority of the trajectories. This seemed to be connected to the favorable pattern of release of the ligand, which lost the interaction with Cys83 and favored the interaction with Asp86 toward the glycine-rich loop.

Proceeding further, at a distance of 11 Å, two contrasting binding patterns emerged (see points E and F in Figures 7 and 9). These were characterized not only by a distinct pattern in the ligand–protein interactions but also by a sizable and opposite rearrangement of the hinge region of the protein.

Finally, while moving toward the solvent, several residual possible interactions were still detectable and sizable (see point G in Figures 7 and 9).

From a drug design perspective, it is potentially important to identify the outliers using local mechanical response analysis.

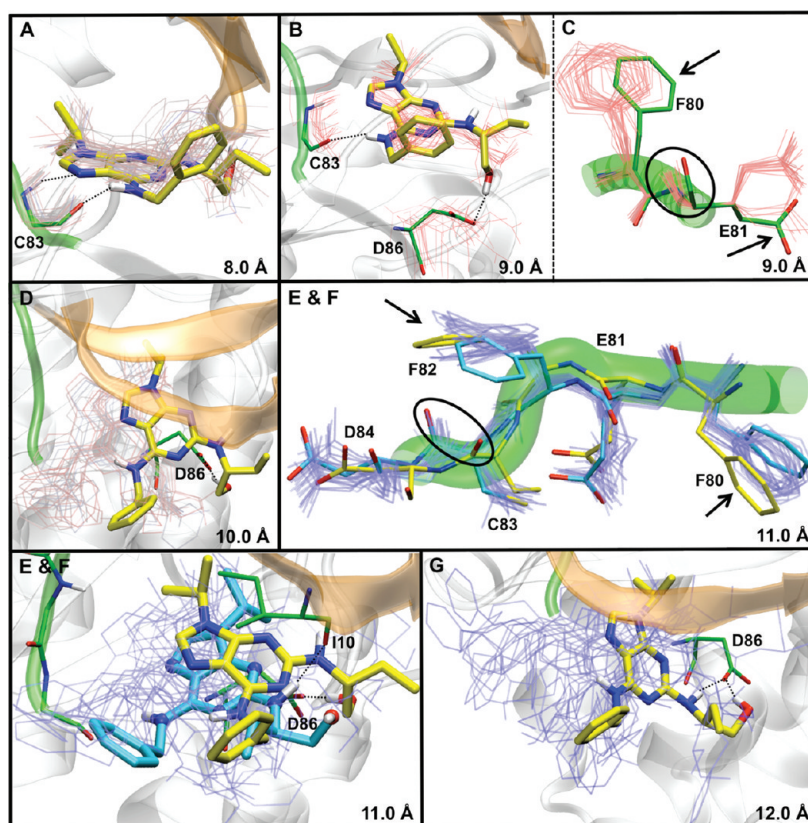


Figure 9. Structural features of the configurations corresponding to the outliers in LMR. The structures highlighted with letters correspond to the circles in Figure 7.

This is because it allows for: (i) the identification of the most relevant structural features responsible for ligand recognition and binding; and (ii) the identification of the key interactions that can be exploited to address chemical modifications aimed at improving the ligand affinity for both its biological counterpart and residence time.¹¹ In practice, this could be achieved by selectively suppressing the low LMR pathways and enhancing the high LMR pathways through specific modifications on the scaffold of the drug.

A crucial aspect of this work is that MDS representations of LMR allow us to group together trajectories that display an analogous mechanical response along the unbinding pathway. This is of great help in focusing the structural analysis on a limited number of representative configurations, which can be crucial when large statistics of trajectories are produced. Additionally, the analysis of outliers in LMR can further strengthen the hypothesis of specific unbinding mechanisms, thus saving a large amount of work in terms of human time. Although the robustness of the results can be improved by increasing the statistics, we have here shown that even a limited statistics of unbinding events can lead to important structural insights.

Moreover, while for this case the use of one single MSD dimension was sufficient, additional dimensions may provide an a further source of flexibility helping to track different exit routes displaying a different mechanical signature.

CONCLUSIONS

In this work, we developed a postprocessing tool for out-of-equilibrium steered MD simulations to identify distinctive

structural features of particular relevance for a specific biological event. Steered MD is nowadays becoming very attractive. Many MD programs now include native routines to perform it.^{36,62} It can also be made available through external plugins.³⁸ One of the major limitations to a wide application of steered MD for proper free energy difference calculations is related to the applicability of the Jarzynski equation, which requires huge computational resources for realistic cases. This is prohibitive in the drug design field, where a major trade-off between speed and accuracy is usually required. We have here shown that very useful structural information can be retrieved from a relatively small statistics of pulling trajectories obtained in a non-quasistatic regime. In particular, in the alanine dipeptide case study, the structural features observed and the topological differences in paths have a clear correspondence in the mechanical response of the system to the external pulling force. The conformational transition of the alanine dipeptide is relevant because of the following points: (i) alanine dipeptide can exist in two well-defined energy minima; (ii) three different paths connect the two minima; (iii) the transitions among minima are usually affected by a very low dissipative work; and (iv) the representation of the conformational transition in the local mechanical space can be straightly converted into a transition in the Cartesian space. Remarkably, while points (i) and (ii) also apply directly to a ligand–protein unbinding, (iii) and (iv) do not. In fact, the ligand unbinding from a protein via steered MD experiences a non-negligible dissipative work, and local mechanical space cannot be directly transformed into a Cartesian space. This is because the protein atoms involved in the unbinding change along the reactive path. Despite this, local

mechanical response can still be very informative from a structural standpoint, as we have here demonstrated by investigating biological systems of increasing complexity. Our second case study was the unbinding of (*R*)-roscovitine from CDK5. Here, the steering coordinate was intentionally chosen to be of general applicability, with the consequent increase in the dissipative work. Nevertheless, even in this regime, many structural details could be obtained. Additionally, from a drug design standpoint, such an approach could also be used in discriminating the understanding of the energetic relationships among alternative docking poses. By adding the nontrivial step of the mechanical response pattern analysis, it could be possible to discriminate those poses that, bringing completely different pathways, are structurally and energetically more or less stable than others.

In conclusion, we have shown that the analysis of the local mechanical response of the system to a forced unbinding can be very informative with respect to the unbinding process itself. Moreover, from a drug design standpoint, it captures the relevant structural events that can be directly exploited to design novel ligands with a potentially increased affinity for and residence time at the biological counterpart. In addition, a large amount of work in terms of human time is saved. This is because researchers need to analyze only the most promising regions where structural dissimilarities are evidenced by MDS. Although the quality of the results can be improved by increasing the statistics, we have shown here that even limited statistics can lead to important insights, which complement a more computationally demanding accurate free energy calculation. In our opinion, this is remarkably important because, as more computer power becomes available and more extended MD studies become possible,⁶³ new analysis techniques^{29–31} as well as new data selection strategies must be devised to optimize storage usage and human effort by directing the action to a smaller but more relevant set of data drawn from a consistent statistics. Additionally, such approach is not limited to ligand–protein binding problems but can be directly applied in simulations of single-molecule force spectroscopy experiments.⁶⁴

■ ASSOCIATED CONTENT

S **Supporting Information.** Selection of spring constant and pulling velocity for alanine dipeptide; representative sketch of the residues of CDK5 involved in the binding process; Cartesian MDS representation of drug unbinding trajectories; selection of the pulling length in undocking of (*R*)-roscovitine/CDK5; effect of steering on the structure of the CDK5; enhanced picture for LMR representation, and steering experiment on (*R*)-roscovitine/CDK5 system with constrained residues. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: davide.branduardi@gmail.com; andrea.cavalli@iit.it.

■ ACKNOWLEDGMENT

We thank the IIT Platform “Integrated Multiscale Computational Technology”, the supercomputer center CINECA (Casalecchio di Reno–BO) and the DEISA Consortium (www.deisa.eu), cofunded through the EU FP6 project RI-031513 and the FP7 project RI-222919, for support within the DEISA Extreme Computing

Initiative. We thank Grace Fox for editing and proofreading the manuscript.

■ REFERENCES

- (1) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J. Chem. Phys.* **1988**, *89*, 3742–3746.
- (2) Gilson, M.; Given, J.; Bush, B.; McCammon, J. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997**, *72*, 1047–1069.
- (3) Hamelberg, D.; McCammon, J. A. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
- (4) Deng, Y.; Roux, B. Calculation of standard binding free energies: aromatic molecules in the T4 Lysozyme L99A mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.
- (5) Wang, J.; Deng, Y.; Roux, B. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **2006**, *91*, 2798–2814.
- (6) Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.* **2009**, *394*, 747–763.
- (7) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
- (8) Mobley, D. L.; Chodera, J. D.; Dill, K. A. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.* **2006**, *125*, 084902–084916.
- (9) Mobley, D. L.; Chodera, J. D.; Dill, K. A. Confine-and-release method: obtaining correct binding free energies in the presence of protein conformational change. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.
- (10) Limongelli, V.; Bonomi, M.; Marinelli, L.; Gervasio, F. L.; Cavalli, A.; Novellino, E.; Parrinello, M. Molecular basis of cyclooxygenase enzymes (COXs) selective inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 5411–5416.
- (11) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discovery* **2006**, *5*, 730–739.
- (12) Swinney, D. C. Biochemical mechanisms of drug action: what does it take for success? *Nat. Rev. Drug Discovery* **2004**, *3*, 801–808.
- (13) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10184–10189.
- (14) Bui, J. M.; Henchman, R. H.; McCammon, J. A. The dynamics of ligand barrier crossing inside the acetylcholinesterase gorge. *Biophys. J.* **2003**, *85*, 2267–2272.
- (15) Roux, B. The calculation of the potential of mean force using computer-simulations. *Comput. Phys. Commun.* **1995**, *91*, 275–282.
- (16) Laio, A.; Parrinello, M. Escaping Free Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *20*, 12562–12566.
- (17) Laio, A.; Gervasio, F. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71*, 126601.
- (18) Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946–5961.
- (19) Masetti, M.; Cavalli, A.; Recanatini, M.; Gervasio, F. L. Exploring complex protein–ligand recognition mechanisms with coarse metadynamics. *J. Phys. Chem. B* **2009**, *113*, 4807–4816.
- (20) Bayas, M. V.; Schulten, K.; Leckband, D. Forced detachment of the CD2-CD58 complex. *Biophys. J.* **2003**, *84*, 2223–2233.
- (21) Gao, M.; Wilmanns, M.; Schulten, K. Steered molecular dynamics studies of titin I1 domain unfolding. *Biophys. J.* **2002**, *83*, 3435–3445.

- (22) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys. J.* **1997**, *72*, 1568–1581.
- (23) Krammer, A.; Lu, H.; Isralewitz, B.; Schulten, K.; Vogel, V. Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1351–1356.
- (24) Colizzi, F.; Perozzo, R.; Scapozza, L.; Recanatini, M.; Cavalli, A. Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *J. Am. Chem. Soc.* **2010**, *132*, 7361–7371.
- (25) Chen, L. Y. Exploring the free-energy landscapes of biological systems with steered molecular dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6176–6183.
- (26) Ozer, G.; Valeev, E. F.; Quirk, S.; Hernandez, R. Adaptive steered molecular dynamics of the long-distance unfolding of neuropeptide Y. *J. Chem. Theory Comput.* **2010**, *6*, 3026–3038.
- (27) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690.
- (28) Xiong, H.; Crespo, A.; Marti, M.; Estrin, D.; Roitberg, A. E. Free energy calculations with non-equilibrium methods: applications of the Jarzynski relationship. *Theor. Chem. Acc.* **2006**, *116*, 338–346.
- (29) Rajan, A.; Freddolino, P. L.; Schulten, K. Going beyond clustering in MD trajectory analysis: an application to villin headpiece folding. *PLoS ONE* **2010**, *5*, e9890.
- (30) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. Wordom: a program for efficient analysis of molecular dynamics simulations. *Bioinformatics* **2007**, *23*, 2625–2627.
- (31) Wriggers, W.; Stafford, K. A.; Shan, Y.; Piana, S.; Maragakis, P.; Lindorff-Larsen, K.; Miller, P. J.; Gullingsrud, J.; Rendleman, C. A.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Automated event detection and activity monitoring in long molecular dynamics simulations. *J. Chem. Theory Comput.* **2009**, *5*, 2595–2605.
- (32) Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling: theory and applications*; 2nd ed.; Springer-Verlag: New York, 2005.
- (33) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (34) Crooks, G. E. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *J. Stat. Phys.* **1998**, *90*, 1481–1487.
- (35) Kearsley, S. K. On the orthogonal transformation used for structural comparison. *Acta Cryst. A* **1989**, *45*, 208–210.
- (36) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (37) Paterlini, M. G.; Ferguson, D. M. Constant temperature simulations using the Langevin equation with velocity Verlet integration. *Chem. Phys.* **1998**, *236*, 243–252.
- (38) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.
- (39) MacKerell, A. D.; Bashford, D.; Bellott, Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (40) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (41) Mapelli, M.; Massimiliano, L.; Crovace, C.; Seeliger, M. A.; Tsai, L. H.; Meijer, L.; Musacchio, A. Mechanism of CDK5/p25 binding by CDK inhibitors. *J. Med. Chem.* **2005**, *48*, 671–679.
- (42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725.
- (43) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (44) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (46) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
- (47) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (48) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (49) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125*, 24106.
- (50) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 054103.
- (51) Bolhuis, P. G.; Dellago, C.; Chandler, D. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877–5882.
- (52) Apostolakis, J.; Ferrara, P.; Caflisch, A. Calculation of conformational transitions and barriers in solvated systems: application to the alanine dipeptide in water. *J. Chem. Phys.* **1999**, *110*, 2099–2108.
- (53) Ferrara, P.; Apostolakis, J.; Caflisch, A. Targeted molecular dynamics simulations of protein unfolding. *J. Phys. Chem. B* **2000**, *104*, 4511–4518.
- (54) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95–99.
- (55) Bibb, J. A. Role of Cdk5 in neuronal signaling, plasticity, and drug abuse. *Neurosignals* **2003**, *12*, 191–199.
- (56) Cheng, K.; Ip, N. Y. Cdk5: a new player at synapses. *Neurosignals* **2003**, *12*, 180–190.
- (57) Dhavan, R.; Tsai, L. H. A decade of CDK5. *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 749–759.
- (58) Gupta, A.; Tsai, L. H. Cyclin-dependent kinase 5 and neuronal migration in the neocortex. *Neurosignals* **2003**, *12*, 173–179.
- (59) Lau, L. F.; Ahljianian, M. K. Role of cdk5 in the pathogenesis of Alzheimer's disease. *Neurosignals* **2003**, *12*, 209–214.
- (60) Nguyen, M. D.; Julien, J. P. Cyclin-dependent kinase 5 in amyotrophic lateral sclerosis. *Neurosignals* **2003**, *12*, 215–220.
- (61) Smith, P. D.; Crocker, S. J.; Jackson-Lewis, V.; Jordan-Sciutto, K. L.; Hayley, S.; Mount, M. P.; O'Hare, M. J.; Callaghan, S.; Slack, R. S.; Przedborski, S.; Anisman, H.; Park, D. S. Cyclin-dependent kinase 5 is a mediator of dopaminergic neuron loss in a mouse model of Parkinson's disease. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13650–13655.
- (62) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.
- (63) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341–346.
- (64) Irback, A.; Mitternacht, S.; Mohanty, S. Dissecting the mechanical unfolding of ubiquitin. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13427–13432.
- (65) Stierand, K.; Maab, P. C.; Rarey, M. Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics* **2006**, *22*, 1710–1716.

GROMOS++ Software for the Analysis of Biomolecular Simulation Trajectories

Andreas P. Eichenberger,[†] Jane R. Allison,[†] Jožica Dolenc,^{†,||} Daan P. Geerke,[§] Bruno A. C. Horta,[†] Katharina Meier,[†] Chris Oostenbrink,[‡] Nathan Schmid,[†] Denise Steiner,[†] Dongqi Wang,[†] and Wilfred F. van Gunsteren^{*,†}

[†]Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH, Zürich, Switzerland

[‡]Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Vienna, Austria

[§]Division of Molecular and Computational Toxicology, Free University Amsterdam, Amsterdam, The Netherlands

^{||}Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, Slovenia

ABSTRACT: GROMOS++ is a set of C++ programs for pre- and postprocessing of molecular dynamics simulation trajectories and as such is part of the **GRO**ningen **MO**lecular **S**imulation software for (bio)molecular simulation. It contains more than 70 programs that can be used to prepare data for the production of molecular simulation trajectories and to analyze these. These programs are reviewed and the various structural, dynamic, and thermodynamic quantities that can be analyzed using time series, correlation functions, and distributions are described together with technical aspects of their implementation in GROMOS. A few examples of the use of GROMOS++ for the analysis of MD trajectories are given. A full list of all GROMOS++ programs, together with an indication of their capabilities, is given in the Appendix.

1. INTRODUCTION

Over the past decades, classical simulation of the dynamics of (bio)molecules in aqueous solution, in crystalline form, or embedded in a lipid membrane has found widespread use in physicochemical, biochemical, and molecular biological research.^{1–5} This has become possible through the availability of a number of general software packages for (bio)molecular simulation, such as AMBER,⁶ CHARMM,⁷ Desmond,⁸ GROMACS,⁹ GROMOS,¹⁰ IMPACT,¹¹ MOLARIS,^{12,13} NAMD,¹⁴ and TINKER,¹⁵ and the development of (bio)molecular force fields,¹⁶ e.g., AMBER,¹⁷ CHARMM,^{18–25} ECEPP,²⁶ ENCAD,²⁷ CFF,²⁸ GROMOS,^{29–35} and OPLS.^{36,37} These simulation software packages generally contain a number of functions that can be used to analyze the atomic trajectories produced in a molecular dynamics (MD) simulation in terms of time series, correlation functions, or trajectory averages of quantities of interest. The set of different quantities that can be analyzed and the computational procedures involved differ between the various simulation packages and are generally scarcely described in the literature. This renders a great deal of research articles irreproducible because the details of the calculations cannot be recovered. Here we present the variety of functions to analyze MD (bio)molecular simulation trajectories that are part of the GROMOS software for (bio)molecular simulation.^{10,38–42}

The GROMOS software is written in C++ and comprises two major parts: (1) MD++, the programs that can be used to perform energy minimization, MD, or stochastic dynamics (SD) simulations, and (2) GROMOS++, the programs that can be used for preprocessing (bio)molecular data prior to a simulation and for postprocessing the trajectories, coordinates, velocities, energies, etc. produced by an MD simulation. The postprocessing functions, in particular, could also be used to analyze (bio)molecular

trajectories generated using (bio)molecular simulation software other than GROMOS. Therefore, we only briefly review the preprocessing capabilities of GROMOS++ and describe the postprocessing analysis functions in more detail.

The architecture, implementation, and parallelization of the GROMOS software is reported elsewhere,³⁹ while the analysis functions that can be used for the analysis of NMR spectroscopic and X-ray and neutron diffraction data are described in ref 42 in conjunction with their use in MD simulation based on such data. Since GROMOS++ is written in an object-oriented manner, new functionality is easily implemented.

2. TRAJECTORY QUANTITIES AND BASIC TYPES OF ANALYSIS

A simulation trajectory, which allows the calculation of many different properties of the simulated system, is primary data for a computer simulation scientist. Although the simulation programs of MD++ perform the main task of generating such simulation trajectories, an appropriate setup of a simulation and analysis of the trajectories are just as important as the simulation itself.

In this section the preprocessing of data required by a molecular simulation is outlined, followed by a discussion of the different types of simulation trajectories that are used for analysis in combination with the GROMOS++ programs. A description of the different types of analysis is given, including technicalities such as the treatment of periodic boundary conditions or the atom, vector and property specifiers implemented in GROMOS++.

Received: May 31, 2011

Published: August 22, 2011

Table 1. Preparation of a Molecular System for an MD Simulation

program	action
make_top	build a molecular topology file holding all the parameters and specifications that characterize the molecular system
com_top	
pdb2g96	converts solute coordinates in Protein Data Bank (PDB) format to GROMOS format
gch	generates hydrogen coordinates based on geometric criteria
MD++	minimizes the intermolecular solute energy to remove possible strain from the solute
ran_box	generates the coordinates for a box with solvent molecules in GROMOS format
sim_box	puts the solute into a solvent box of appropriate size and removes all solvent molecules that show a given spatial overlap with solute atoms
MD++	minimizes the system energy while the solute is kept positionally restrained to remove high-energy intermolecular contacts
ion	replaces solvent molecules by ions in order to neutralize the Coulomb charge of the solute or attain a given salt content

There are three questions to be answered at the beginning of each molecular simulation: (1) What is the chemical structure of the system to be simulated, i.e. the number and types of elements, atoms, particles, and their connectivity? (2) Which force field, containing the parameters to model the interactions in the system based on physical theories and approximations, is to be used? (3) How is the interface with the surroundings of the molecular system to be modeled? The answers to the first two questions allow a system to be set up for MD simulation using the sequence of steps specified in Table 1. A short description of the functionality of each program is given in Table A1 of the Appendix. A more detailed description can be found in the GROMOS manual or the electronic code documentation of GROMOS++.

2.1. Types of Trajectory Information Handled by GROMOS++. A simulation trajectory is completely characterized by the time series of the particle coordinates and velocities, from which a variety of quantities can be calculated. However, it is efficient to also have the possibility to save forces on the particles or energetic quantities of particular sets of atoms during an MD simulation, because the calculation of these types of quantities is computationally expensive and should not have to be repeated while postprocessing MD simulation data. Depending on the input parameters, the GROMOS MD++ simulation program regularly writes specific information about the system to the corresponding trajectory file. The time interval between saved configurations should be small enough to obtain sufficient configurations for averaging, but it is generally chosen to be much larger than the MD time step in order to avoid storage of correlated data, unless, of course, the short-time correlations happen to be of interest. Depending on the desired quantity to be calculated, a postprocessing program of GROMOS++ reads one or multiple simulation trajectory files of a specific type to compute various other, more complex quantities. GROMOS++ provides more than 70 programs (see the Appendix) ready to be used for pre- and postprocessing of molecular simulations.

The following is an overview of four different types of GROMOS simulation trajectories handled by GROMOS++. The focus is on the information content of each simulation trajectory type and not on its use for analysis. The latter should be clear after reading the sections below.

Positions and Velocities. These simulation trajectory files contain—besides the time information, i.e., current integration time step and simulated time—the three-dimensional, Cartesian atom positions and atom velocities. Additionally, the box size is indicated for every configuration of the trajectory. Information such as atom names and connectivities is generally omitted but can easily be recovered with help of the corresponding molecular topology file based on the GROMOS rule that the sequence of atoms is identical in both files.

Forces. The force trajectory file is similar to the position and velocity trajectory files. It stores atomic Cartesian forces including those induced by application of constraints on the system, typically using SHAKE.⁴³

Energies. Besides the time step information, energy trajectories consist of two main parts, the first being the total energy of the system and its components: kinetic and potential energies, bonded and nonbonded contributions, lattice-sum terms, self-polarization contributions when using a polarizable force field, and special energy terms, e.g., those arising from solvent accessible surface area (SASA) implicit solvent simulations,⁴⁴ from diverse restraining functions, or from enveloping distribution sampling (EDS) simulations.⁴¹ Moreover, it contains kinetic energies for the different sets of atoms that are coupled to a thermal bath as well as the bonded, nonbonded, and special energy contributions of each so-called energy group (predefined in the MD++ input file). The second part of the energy trajectory contains information about the mass, the box dimensions, the temperature, and the pressure of the simulated system. Some properties are stored in a 3×3 tensor format, namely, the pressure, the virial, and the molecular kinetic energy. These data can also be stored as block averages.

Free Energy Data. A free energy trajectory contains the same terms of the Hamiltonian as the energy trajectory described above, but as derivatives with respect to the coupling parameter λ . Such an approach is usually credited to Kirkwood,⁴⁵ who used it to derive expressions for the chemical potential of components of mixtures. The λ -dependence is useful for the calculation of relative free energy differences, ΔF , e.g. via thermodynamic integration (TI),

$$\Delta F = F(\lambda_B) - F(\lambda_A) = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (1)$$

where the Hamiltonian H describes different systems or states of a system as a function of λ .

2.2. Types of Analysis. GROMOS++ contains a number of programs that can be used in different combinations for the calculation of a variety of quantities. The majority of these programs belong to one of the following groups, depending on the type of analysis they perform: time series, distributions, or time correlation functions.

Time series can be calculated for a variety of scalar or vector quantities $Q(t)$ or $\mathbf{Q}(t)$, respectively, which are defined in terms of Cartesian atomic coordinates or velocities. Examples of such quantities are bond angles or torsional dihedral angles or vectors defined by atoms in the molecule, inner and outer products of such vectors, etc. GROMOS++ is able to write time series for such quantities based on position or velocity trajectories. By using energy or free energy trajectories it is possible to generate

time series of temperatures, densities, or energetic properties in a straightforward manner.

Distributions of a quantity Q can be calculated using GROMOS++. Normalization allows for the computation of probabilities $P(Q)$, while the average $\langle Q \rangle_t$ and the mean-square fluctuation ΔQ^2 are accessible as the first and second moment of the distribution,

$$\Delta Q^2 = \langle (Q - \langle Q \rangle_t)^2 \rangle_t \quad (2)$$

Time correlation functions $C_Q(t)$ or $C_{\mathbf{Q}}(t)$ are defined as

$$\begin{aligned} C_Q(t) &= \langle f(Q_i(t'), Q_j(t' + t)) \rangle_{t'} \\ &= (t_{\text{MD}} - t)^{-1} \int_0^{t_{\text{MD}} - t} f(Q_i(t'), Q_j(t' + t)) dt' \end{aligned} \quad (3)$$

The function f may be a simple multiplication of the two quantities $Q_i(t')$ and $Q_j(t' + t)$, but GROMOS++ allows a user-specified definition of this function. $C_{\mathbf{Q}}(t)$ is defined in the same way, but depends on the vector quantities $\mathbf{Q}_i(t)$ and $\mathbf{Q}_j(t' + t)$ instead of $Q_i(t')$ and $Q_j(t' + t)$. If $i = j$, the correlation function is an autocorrelation function, while for other cases it is a cross-correlation function. The calculation of Q_i and Q_j from a trajectory is usually done for N_t discrete, equally spaced time points $t_n = n\Delta t$ with $n = 0, 1, \dots, N_t - 1$. The discrete equivalent of eq 3 is then

$$C_Q(n\Delta t) = (N_t - n)^{-1} \sum_{k=0}^{N_t - n - 1} f(Q_i(k\Delta t), Q_j((k + n)\Delta t)) \quad (4)$$

If $f(Q_i(t'), Q_j(t' + t)) = Q_i(t')Q_j(t' + t)$, GROMOS++ makes use of fast Fourier transforms instead of the more time-consuming summation algorithm.

2.3. Superposition of Molecular Structures. Molecules tumble in space during a molecular simulation. This is exactly what they should do when in the liquid phase. However, if this overall motion of the solute is of no interest, GROMOS++ programs can perform an alignment of the configuration of one or multiple molecules against a reference configuration or structure, e.g. the programs `rmsd` and `rmsf`; see the Appendix. This alignment is carried out by first superimposing the centers of mass of a defined set of atoms in each configuration, followed by a rotational fit with respect to this subset of molecular atom positions. This superposition of pairs of structures is done automatically when running the programs for which it is required, while the atoms and the reference structure to be used for the superpositioning can be specified by the user.

2.4. Spatial Boundary Conditions and Gathering. Cartesian position trajectories generated using (periodic) boundary conditions with the following box shapes can be handled by GROMOS++: vacuum, rectangular, truncated octahedron, and triclinic. Due to the periodicity of the nonvacuum boundary conditions, the trajectory coordinates of molecules or groups of atoms may be such that covalent bonds are broken; i.e., the nearest image of an atom may not be the one saved in the trajectory file. Before calculating interatomic quantities such as bond lengths, bond angles, torsional dihedral angles, etc., GROMOS++ offers the possibility to select the nearest-image position with respect to a reference position, i.e., to gather a trajectory before the analysis. The different gathering methods that may be chosen by the user are listed in Table 2.

Table 2. Handling of Periodic Boundary Conditions

method	action
nog	no gathering
glist	gathering with respect to a list of atom pairs
gtime	gathering with respect to the previous configuration of the trajectory
gref	gathering with respect to a reference structure or configuration of atoms
gltime	gather the first configuration of the trajectory based on a list of atoms and the following configurations with respect to the previous configuration
grtime	gather the first configuration of the trajectory based on a reference configuration and the following configurations with respect to the previous configuration
gbond	gathering based on the bond connectivities of the solute

2.5. Atom, Vector, and Property Specifiers. The more specific the analysis of a simulation, the higher the requirements with respect to the flexibility of the corresponding analysis tool. GROMOS++ makes use of three specifiers to define the quantity to be calculated as precisely and compactly as possible while a high level of flexibility is kept. All three specifiers are passed to the program as input parameters.

Atom specifiers offer a flexible way to access specific atoms of a system. It is even possible to specify atoms that are not present in the trajectory or molecular topology file, so-called virtual atoms, e.g. hydrogen atoms of a united CH_x atom, or common properties of a group of atoms, such as the center of geometry or the center of mass. Atom specifiers are of the format

`<molecule>:<atoms1>[,<atoms2>, ..., <atomsN>]`

where `<molecule>` is the molecule number and `<atoms>` is one or more atoms defined in one of the following ways: the atom number within the molecule `<molecule>`, the atom name, or the residue and atom, both specified either by its number or name. Virtual atoms are accessed via their type and the atoms needed to construct the defined virtual atom type. For example a (virtual) aromatic hydrogen atom position is generated by the three neighboring CH_1 united atom positions, based on geometric criteria.⁴⁶

A more complicated selection of atoms is accessed using the options “?”, `minus(<atom specifier>)`, and `not(<atom specifier>)`. The wild card “?” is used to specify groups of atoms with similar atom names, e.g. all carbon atoms (CA, CB, CG,...) of a protein:

`1:C?`

`minus(<atom specifier>)` and `not(<atom specifier>)` both specify atoms not to be selected. The option `minus(<atom specifier>)` allows atoms to be added later on, while `not(<atom specifier>)` definitely removes the specified atoms from the selection. The following three examples all specify all CA atoms of odd numbered residues of a 10 amino acid peptide:

`1:res(1,3,5,7,9:CA)`

`1:CA minus(1:res(2,4,6,8,10:CA))`

`not(1:res(2,4,6,8,10:CA)) 1:CA`

Table 3. Property Specifiers

specifier	property
d	the distance between two atoms
a	the angle defined by three atoms
t	a torsional dihedral angle
hb	the presence of hydrogen bonds
st	the stacking between two groups of atoms
o	the order between two vectors
op	the order parameter
pr	a pseudorotation
pa	the pucker amplitude
expr	a quantity defined by an expression property

It is also possible to read atom specifiers from a file, e.g. written by the program `atominfo`, using the atom specifier `file(<file name>)`.

Property Specifiers. Some GROMOS++ programs are able to read a property specifier, a helpful tool to describe the quantity to be calculated. The syntax for the property specifier is

$$\langle \text{type} \rangle \% \langle \text{arg} \rangle$$

with $\langle \text{type} \rangle$ defining the property and $\langle \text{arg} \rangle$ an atom or vector specifier providing the necessary information to calculate the desired quantity. Vector specifiers are described below. The different types of property specifiers are given in Table 3. The expression type, `expr`, is a specifier itself with the following syntax:

$$\text{expr} \% \langle \text{f1} \rangle (\langle \text{args1} \rangle) \langle \text{op} \rangle \langle \text{f2} \rangle (\langle \text{args2} \rangle)$$

where $\langle \text{op} \rangle$ is an arithmetic or logical operator, while $\langle \text{f1} \rangle$ and $\langle \text{f2} \rangle$ are functions with the corresponding scalar or vector arguments, $\langle \text{arg1} \rangle$ or $\langle \text{arg2} \rangle$, depending on the nature of the function. The following functions are supported: `sin`, `cos`, `tan`, `asin`, `acos`, `atan`, `abs` (the norm of a scalar or vector), `sqrt`, `abs2` (the squared norm of a vector), `dot`, `cross`, or `ni` (nearest image).

Vector Specifiers. Property and expression specifiers may be defined as a function of one or more vectors. In GROMOS++, a vector can be specified by its three-dimensional Cartesian coordinates

$$\text{cart}(\langle x \rangle, \langle y \rangle, \langle z \rangle)$$

or in polar coordinates

$$\text{polar}(\langle r \rangle, \langle \alpha \rangle, \langle \beta \rangle)$$

with r being the length of the vector and α, β the two angles to orient the vector in space. In addition, the Cartesian coordinates of atoms may be used to specify a vector.

$$\text{atom}(\langle \text{atom specifier} \rangle)$$

refers to the coordinates of the atom defined by an atom specifier. If the atom specifier holds two atoms, the vector is specified pointing from the first to the second atom.

2.6. Input/Output Formats. All GROMOS++ programs are called from the command line, taking the necessary additional information as command line arguments. A usual program call looks like

$$\text{program} @ \langle \text{flag1} \rangle \langle \text{arg1} \rangle [@ \langle \text{flag2} \rangle \langle \text{arg2} \rangle \\ \dots @ \langle \text{flagN} \rangle \langle \text{argN} \rangle]$$

The different arguments may also be collected in a single input file. The program call then looks like

$$\text{program} @ \text{f file}$$

Since GROMOS++ is used for pre- and postprocessing of molecular simulations, its input and output functionality is complex. Input flags and arguments may point to files of various formats (e.g., topology files, simulation trajectories, NOE bound specifications) that are described in more detail in the GROMOS manual. However, GROMOS++ usually writes its output to the standard output, which may be directed into a text file.

The reading and writing of coordinate files (single configuration or trajectories) is often done and is therefore optimized: compressed trajectory files can be read directly and are decompressed while reading. Additionally, the writing of coordinate files may be done in one of the following user specified formats: standard or reduced GROMOS format, PDB format, and AMBER format.

Further, the two argument flags `@inG96` and `@outG96` ensure the compatibility with earlier GROMOS versions that read or write previous file formats.

2.7. Physical Units. Physical constants define the units used in GROMOS and are therefore not hard coded in GROMOS++. If a topology file is given, the physical constants and their units are derived from the following four physical constants, defined in the molecular topology file: Boltzmann's constant, Planck's constant, the electric permittivity of vacuum, and the speed of light. Analyses carried out without the information of a molecular topology file use physical constants that are initialized to defaults while printing a warning.

The GROMOS example files use the following basic units of the standard international system:

length, nm = 10^{-9} m

mass, u = atomic mass unit

time, ps = 10^{-12} s

temperature, K

charge, e = electronic charge = $1.602\,177\,3 \times 10^{-19}$ C

3. OBSERVABLE AND NONOBSERVABLE QUANTITIES

GROMOS++ consists of more than 70 individual programs for many different tasks. To describe them all in detail is beyond the scope of this work. However, we report the calculation of some structural, dynamic, and thermodynamic quantities, covering the practically most relevant aspects of the postprocessing functionalities of GROMOS++. The calculation of experimentally observable and derived quantities is briefly mentioned but described in more detail elsewhere.⁴² A complete list of all current GROMOS++ programs with a short description of each is found in the Appendix (Tables A1 and A2).

3.1. Structural Quantities. `rmsd`, `rmsf`, `rgyr`. These programs are mainly developed for the analysis of structural properties of a peptide or protein, but may also work for other molecules. The atom positional root-mean-square deviation (RMSD) between two molecular structures is calculated according to

$$\text{RMSD}(\mathbf{r}^{N_i}, \mathbf{r}_{\text{ref}}^{N_i}) = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (\mathbf{r}_i - \mathbf{r}_{i,\text{ref}})^2} \quad (5)$$

where $\mathbf{r}^{N_a} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_{N_a})$ indicates the positions of the atoms, N_a the number of atoms considered, \mathbf{r}_i the position of atom i in the first structure, and $\mathbf{r}_{i,\text{ref}}$ the position of atom i in the second, reference structure.

The atom positional root-mean-square fluctuation (RMSF) of an atom i is computed according to

$$\text{RMSF}_i = \sqrt{\frac{1}{N_T} \sum_{t=1}^{N_T} (\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle)^2} \quad (6)$$

where $\langle \mathbf{r}_i \rangle$ is the time-averaged position of atom i and N_T the number of configurations or time frames in the simulation trajectory.

In contrast to the RMSD and RMSF functions, which describe the positional change and mobility of atoms of a molecule, the radius of gyration (R_{gyr}) is a measure of the compactness of the structure. It can be related to light-scattering intensities and is calculated using the definition

$$R_{\text{gyr}} = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (\mathbf{r}_i - \mathbf{R}_{\text{cm}})^2} \quad (7)$$

with

$$\mathbf{R}_{\text{cm}} = \frac{1}{M} \sum_{i=1}^{N_a} m_i \mathbf{r}_i \quad (8)$$

and

$$M = \sum_{i=1}^{N_a} m_i \quad (9)$$

in which \mathbf{r}_i denotes the Cartesian position of atom i , m_i its mass, and N_a the number of considered atoms.

sasa. The solvent-accessible surface area can be calculated as described by Lee and Richerds.⁴⁷ A spherical probe of given radius r is rolled over the surface of the solute. The path traced out by its center is proportional to the solvent-accessible surface area.

In GROMOS, the radii of the heavy atoms are obtained by calculating the minimum energy distance of the interaction between the solute atom and the solvent. This value is reduced by the specified probe radius r to account for the radius of the solvent atom.

Alternatively, program *sasa_hasel*, which contains an implementation of the algorithm described by Hasel et al.,⁴⁸ can be used.

dssp. The detection of secondary structure in a protein is implemented according the rules of Kabsch and Sander.⁴⁹ An overview over the different residues with the percentage assigned to each secondary structure element over the trajectory is printed, while time series for each secondary structure element are saved in separate files. It may occur that one residue is assigned to be part of two secondary structure elements at the same time. In order to avoid ambiguous assignments, the following priority rules are applied: β -sheet/bridge > α -helix > π -helix > 3_{10} -helix > hydrogen-bonded turn > bend.

hbond. The occurrence of two-center and three-center hydrogen bonds defined using geometric criteria can be monitored over a trajectory.⁵⁰ A two-center hydrogen bond is considered to be present if (1) the position of a hydrogen atom H connected to a donor atom D is within $d_{\text{HA}} = 0.25$ nm from that of an acceptor atom A and (2) the D–H–A angle is larger than $A_{\text{DHA}} = 135^\circ$. Occurrences of three-center hydrogen bonds are defined for a

donor atom D, hydrogen atom H, and two acceptor atoms A_1 and A_2 , if

- (i) the distances H– A_1 and H– A_2 are smaller than 0.27 nm,
- (ii) the angles D–H– A_1 and D–H– A_2 are larger than 90° ,
- (iii) the sum of the angles D–H– A_1 , D–H– A_2 , and A_1 –H– A_2 is larger than 340° , and
- (iv) the dihedral angle defined by the planes through the atoms D– A_1 – A_2 and H– A_1 – A_2 is smaller than 15° .

All distance and angle bounds used in the criteria above may be changed by the user. If a reference structure is given, only the hydrogen bonds present in the reference structure are monitored. Otherwise, all intramolecular and/or intermolecular hydrogen bonds between user-specified subsets of atoms in the system are followed. Averages for the monitored distances, angles, and occurrences are printed to the standard output while the corresponding time series are written to a file.

tser and *tcf*. Structural quantities defined by a property specifier, e.g., distances, angles, torsional angles, and more complex, user-specified structural properties to be calculated from the atomic coordinates, can be computed as a time series and/or a (normalized) distribution. The output of the program *tser* may be used for further treatment with the program *tcf*, which computes time correlation functions according to eq 3.

dipole. The molecular electric dipole moment \mathbf{p} can be calculated according to

$$\mathbf{p}^{N_a}(\mathbf{r}^{N_a}) = \sum_{i=1}^{N_a} q_i \mathbf{r}_i \quad (10)$$

where $\mathbf{r}^{N_a} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_{N_a})$ indicates the atom positions, N_a is the number of atoms considered, q_i the charge of atom i , and \mathbf{r}_i its Cartesian position. In the case of a total net charge Q of the N_a atoms, $Q = q_1 + q_2 + \dots + q_{N_a} \neq 0$, the dipole moment is dependent on the origin of the coordinate system. Either a particular origin, i.e., the center of geometry of the N_a atoms,

$$\mathbf{p}^{N_a}(\mathbf{r}^{N_a}) = \sum_{i=1}^{N_a} q_i (\mathbf{r}_i - \mathbf{R}_{\text{cog}}) \quad (11)$$

with

$$\mathbf{R}_{\text{cog}} = \frac{1}{N_a} \sum_{i=1}^{N_a} \mathbf{r}_i \quad (12)$$

can be chosen or a uniform background charge of $-Q/N_a$ is added to each atomic charge to make \mathbf{p}^{N_a} origin independent.

cluster. This program performs a conformational clustering based on a similarity matrix as calculated by the program *rmsdmat*. The clustering algorithm is described by Daura et al.⁵¹ Structures with RMSD values smaller than a user specified cutoff are considered to be structural neighbors. The structure with the highest number of neighbors is considered to be the central member of the cluster of similar structures forming a conformation. After removing all structures belonging to this first cluster, the procedure is repeated to find the second, third, etc. most populated clusters.

A specific structure can be forced to be the central member structure of the first cluster, which can also be the reference structure. The clustering can be performed on a subset of the matrix by specifying the maximum number of structures to consider. This allows for an assessment of the development of the number of clusters over time.

3.2. Dynamic Quantities. *diffus.* This program calculates the diffusion constant D of an atom or of the center-of-geometry of a specified set of atoms. First, the mean square-displacement, $\Delta(t)$, is obtained by averaging over all considered atoms and over multiple time windows,

$$\Delta(t) = \frac{1}{N_a} \sum_{i=1}^{N_a} \langle (\mathbf{r}_i(t + \tau) - \mathbf{r}_i(\tau))^2 \rangle_{\tau \leq t_{av} - t} \quad (13)$$

where \mathbf{r}_i is the position of atom i , N_a the number of atoms considered, and t_{av} the averaging time. According to the Einstein relation, the diffusion constant can be estimated from the slope of $\Delta(t)$

$$D = \lim_{t \rightarrow \infty} \frac{\Delta(t)}{2N_d t} \quad (14)$$

where N_d is the number of dimensions considered.

rot_rel. The rotational relaxation time of a molecule can be estimated from the autocorrelation function of the Legendre polynomials of a molecular vector \mathbf{v} of unit length,

$$C_1(t) = \langle \mathbf{v}(\tau) \cdot \mathbf{v}(\tau + t) \rangle_{\tau} \quad (15)$$

$$C_2(t) = \frac{1}{2} \langle 3\langle \mathbf{v}(\tau) \cdot \mathbf{v}(\tau + t) \rangle_{\tau}^2 - 1 \rangle \quad (16)$$

The program *rot_rel* calculates the first- and second-order Legendre polynomials and the time correlation functions. The output of this program can also be produced by a combination of the programs *tser* and *tcf*.

visco. The bulk and shear viscosities can be calculated using the Einstein relation from the elements of the pressure tensor that are written to an energy trajectory. Let $P_{\alpha\beta}$ be the element of the pressure tensor, and $G_{\alpha\beta}(t)$ the time integral of $P_{\alpha\beta}$,

$$G_{\alpha\beta}(t) = \int_0^t P_{\alpha\beta}(t') dt' \quad (17)$$

The viscosity tensor element $\eta_{\alpha\beta}$, calculated in terms of the integral (eq 17) of the pressure component $P_{\alpha\beta}$, is proportional to the mean-square change of $G_{\alpha\beta}(t)$ in the limit of infinite time,

$$\eta_{\alpha\beta} = \frac{V}{2k_B T} \lim_{t \rightarrow \infty} \frac{d}{dt} \langle (G_{\alpha\beta}(t + \tau) - G_{\alpha\beta}(\tau))^2 \rangle_{\tau \leq t_{av} - t} \quad (18)$$

where V denotes the volume of the (periodic) box, k_B the Boltzmann constant, and T the temperature of the system. For isotropic systems, an estimate of the bulk viscosity can be obtained from the average of the three diagonal components of the pressure tensor,

$$\eta_{\text{bulk}} = \frac{1}{3} (\eta_{xx} + \eta_{yy} + \eta_{zz}) \quad (19)$$

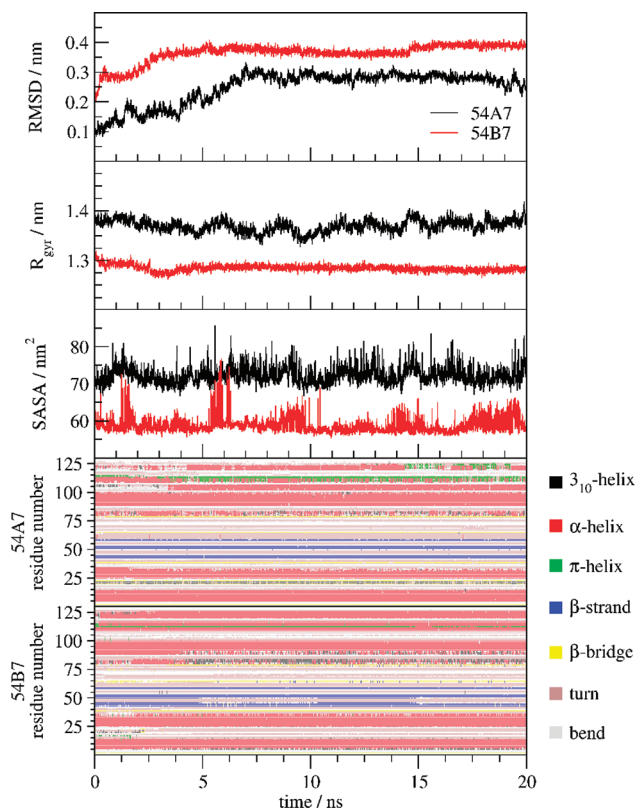


Figure 1. Atom-positional root-mean-square deviation (RMSD) from the initial X-ray structure, radius of gyration (R_{gyr}), solvent accessible surface area (SASA), and secondary structure analysis of two HEWL simulation trajectories using the GROMOS force fields 54A7 (protein in water) and 54B7 (protein in vacuum).³⁵

while the shear viscosity is obtained by averaging the off-diagonal elements,

$$\eta_{\text{shear}} = \frac{1}{3} (\eta_{xy} + \eta_{xz} + \eta_{yz}) \quad (20)$$

eps_field. The static dielectric permittivity, $\epsilon(0)$, of a liquid can be obtained by applying an external electric field during a simulation and measuring the polarization response.⁵² For an external field of strength E_z^{ext} , the permittivity is given by

$$\epsilon(0) = 1 + 4\pi \frac{\langle P_z \rangle_t}{E_z^{\text{ext}}} \quad (21)$$

where P_z is the average polarization of the system in the z -direction and \mathbf{P} is defined as

$$\mathbf{P}(t) = \frac{\mathbf{M}(t)}{V(t)} \quad (22)$$

where \mathbf{M} denotes the total dielectric dipole moment of the system and V the volume of the simulation box.

The external electric field size should be chosen to be small enough to avoid saturation and large enough to induce a significant $\langle P_z \rangle_t$.

3.3. Thermodynamic Quantities. *ene_ana.* This program can calculate time series, time averages, root-mean-square fluctuations, and statistical error estimates for properties contained in an energy or free energy trajectory file, e.g.

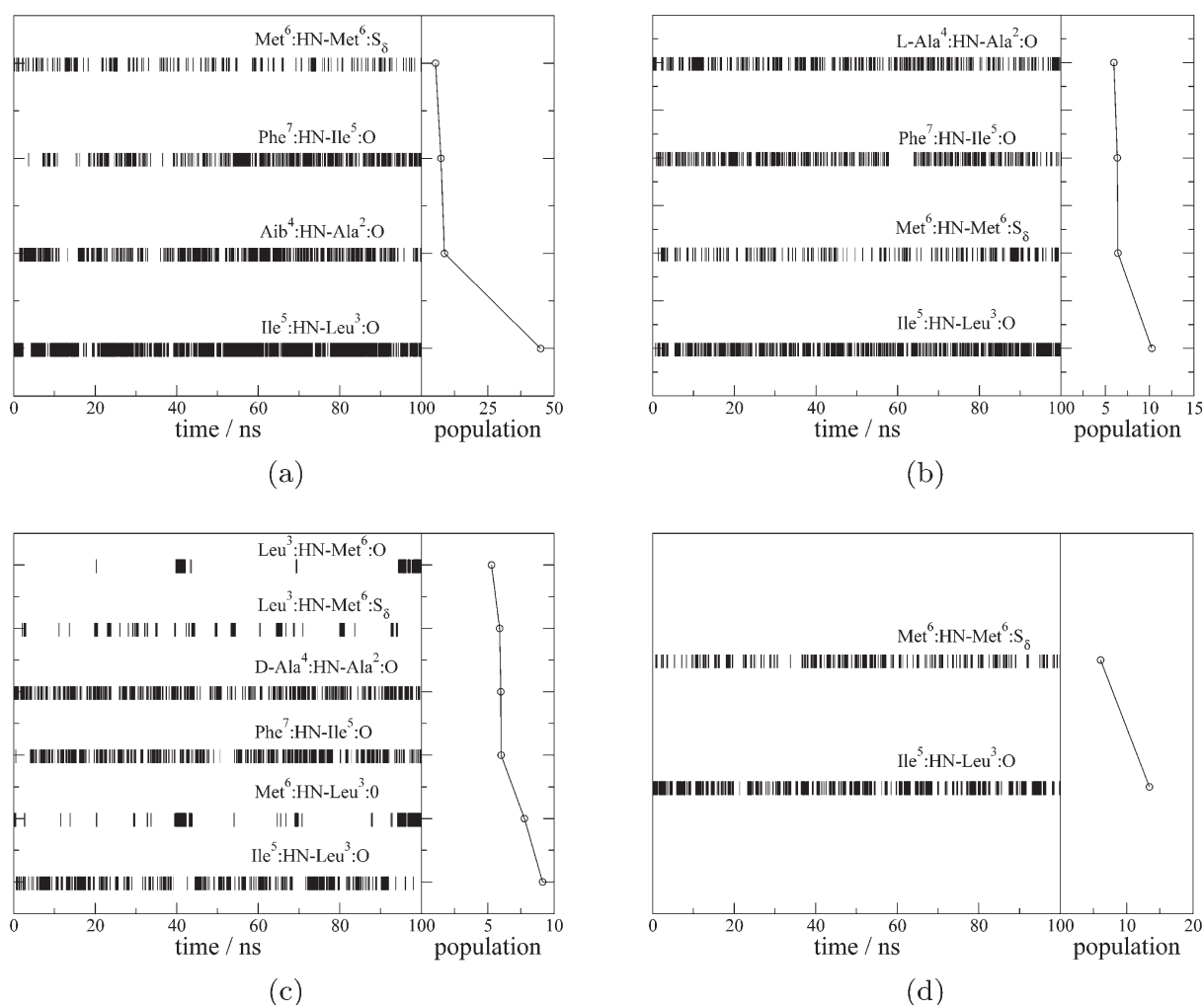


Figure 2. Occurrence of intrasolute hydrogen bonds with a population larger than 5% in the four simulations of the four heptapeptides Val-Ala-Leu-X-Ile-Met-Phe in water: (a) X = Aib, (b) X = L-Ala, (c) X = D-Ala, and (d) X = Gly.⁵⁸

volume, temperature, pressure, different energy terms, entropy, etc. The error estimates are calculated from block averages of different sizes. In combination with a library file, `ene_ana` allows the analysis of quantities that are a function of one or more properties provided by the energy and/or free energy trajectory files. Note that in principle one can teach `ene_ana` to read any trajectory, i.e., also force- or block-averaged energies or positions, by specifying the corresponding block names and block formats in its library file.

Free Energy Differences. The calculation of relative free energies of ligand-protein binding, of solvation for different compounds, and of different conformational states of a (bio)molecule is of considerable interest with regard to an understanding of these processes and to the design or selection of potential inhibitors of enzymes. Since such processes in aqueous solution generally comprise energetic and entropic contributions from many molecular configurations, adequate sampling of the relevant parts of configurational space when calculating ensemble averages is required and can be reached through MD simulations. Most methods to obtain relative free energies require a particular modification of the Hamiltonian in an MD simulation, which leads to artificial forces on the atoms that enhance the sampling.

The implementation in the GROMOS software of the most popular or promising of such methods, i.e. thermodynamic integration, umbrella sampling, local-elevation umbrella sampling, and enveloping distribution sampling, are described elsewhere.⁴¹ Multiple GROMOS++ programs are available for the calculation of free energy differences from simulations with modified Hamiltonians.

Some methods to compute relative free energies only require postprocessing of trajectory data from a standard MD simulation and are therefore mentioned here.

In the Widom particle-insertion method,⁵³ a test or virtual atom is inserted randomly N_{try} times in each configuration of the molecular system and its free energy of “solvation” is then calculated as

$$\Delta F_{\text{solv}} = -k_B T \ln \left[\left\langle \frac{1}{N_{\text{try}}} \sum_{i=1}^{N_{\text{try}}} \exp \left(-\frac{V(\mathbf{r}^N, \mathbf{r}_{\text{test}})}{k_B T} \right) \right\rangle_{\mathbf{r}^N} \right] \quad (23)$$

where $V(\mathbf{r}^N, \mathbf{r}_{\text{test}})$ is the potential energy of the test particle with respect to all atoms in the system and the average is over

all configurations of the N atoms of the system. The program `m_widom` calculates this free energy.

In the one-step perturbation method,^{54,55} the free energy change due to a change of the reference Hamiltonian H_R into a perturbed Hamiltonian H_A is calculated as

$$\begin{aligned} \Delta F_{AR} &= F_A - F_R \\ &= -k_B T \ln \left[\left\langle \exp \left(-\frac{(H_A - H_R)}{k_B T} \right) \right\rangle_R \right] \end{aligned} \quad (24)$$

where the ensemble average is over the simulation based on H_R . Such an ensemble average can be calculated using the program `dg_ener`.

When using a biasing potential energy term $V_{US}(\mathbf{r}^N)$ in the Hamiltonian $H_{bias}(\mathbf{r}^N)$ of the MD simulation,⁵⁶ its influence has to be removed from the ensemble averages. This is done by so-called reweighting of the configurations in the averaging:

$$\langle Q \rangle = \frac{\left\langle Q \exp \left(\frac{V_{US}}{k_B T} \right) \right\rangle_{bias}}{\left\langle \exp \left(\frac{V_{US}}{k_B T} \right) \right\rangle_{bias}} \quad (25)$$

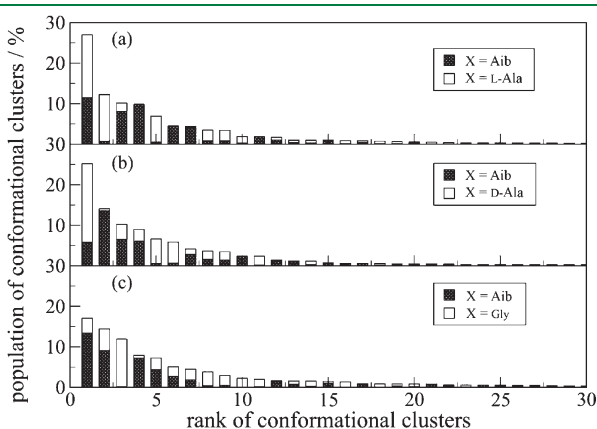


Figure 3. Conformational clustering analysis over the 100 ns trajectories of the four heptapeptides [atom-positional RMSD within 0.08 nm for backbone N, C(β), C(α), and C-atoms of residues 2–6].⁵⁸ In each panel the population of clusters observed in two joint 100 ns trajectories are shown, with the number of configurations originating from the X = Aib peptide indicated in gray.

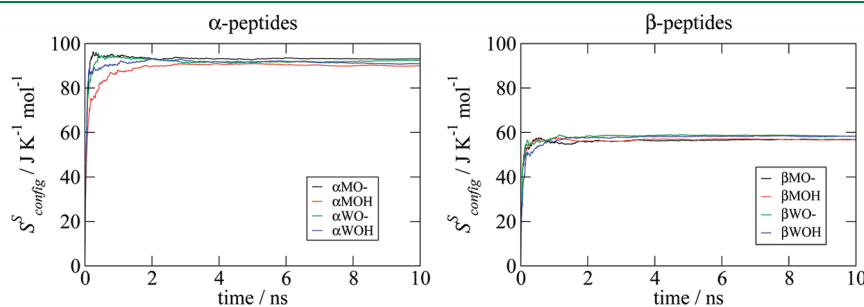


Figure 4. Buildup of the Schlitter entropy S_{config}^S , eq 26, as function of time. The Schlitter entropy was calculated for the C_α atoms during 10 ns MD simulations of various 24 backbone atom α - and β -peptides in solution as labeled (α MO- = α -peptide, methanol, O⁻ terminal group).⁵⁹ In each case, the structures were first aligned according to the positions of their C_α atoms.

The ensemble average $\langle \dots \rangle$ for an unbiased Hamiltonian H is expressed in terms of two ensemble averages for the biased Hamiltonian $H_{bias} = H + V_{US}$. Reweighting can be performed using the program `reweight`.

The configurational entropy of a solute molecule can be estimated using Schlitter's heuristic formula,⁵⁷ which gives an upper bound for the true entropy

$$S_{true} \leq S = \frac{1}{2} k_B \ln \left[\det \left(1 + k_B T \left(\frac{e}{\hbar} \right)^2 \mathbf{M} \boldsymbol{\sigma} \right) \right] \quad (26)$$

where \hbar is Planck's constant divided by 2π , e is Euler's number, \mathbf{M} is the diagonal mass matrix holding on the diagonal the masses belonging to the $3N_a$ Cartesian degrees of freedom, and $\boldsymbol{\sigma}$ is the covariance matrix of the positional fluctuations of these degrees of freedom. The elements of $\boldsymbol{\sigma}$ are

$$\sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \quad (27)$$

with x_i being the Cartesian coordinates of the N_a atoms considered for the entropy calculation after least-squares fitting of the position of a given subset of atoms of the coordinate trajectory. The program `solute_entropy` calculates Schlitter entropy and the quasi-harmonic entropy from molecular coordinate trajectories.

3.4. Comparison to Experimentally Observable and Derived Quantities. GROMOS++ is able to calculate a variety of experimentally observable quantities as well as quantities that are derived from experimental data, i.e., NOE intensities or atom–atom distance bounds, 3J -coupling constants, residual dipolar couplings (RDCs), order parameters derived from NMR experiments, small- and wide-angle X-ray scattering (SAXS/WAXS) intensities, and total neutron scattering intensities for liquids. A detailed description of the implementation of such quantities in the GROMOS software and their use in protein refinement, including examples, is described by Schmid et al.⁴²

4. EXAMPLES

Here we present a few examples of the use of GROMOS++ programs to analyze MD simulation trajectories, taken from previous work.

4.1. Global Structural Properties as Function of Time. In Figure 1, the backbone atom-positional RMSD from the initial X-ray-derived structure for two MD simulations of hen egg

white lysozyme (HEWL) is given as a function of time together with the radius of gyration, the solvent accessible surface area, and the occurrence of major secondary structure elements. The trajectory of the simulation in vacuo clearly deviates more from the X-ray structure than that of the simulation in water, illustrating the effect of the solvent on protein structure. The setup of these simulations is described elsewhere.³⁵

4.2. Hydrogen Bond Analysis as Function of Time. In Figure 2 the intrapeptide hydrogen bonding of four seven-residue peptides in methanol as observed in 100 ns MD simulations of their folding equilibria is shown. The four peptides with sequence Val-Ala-Leu-X-Ile-Met-Phe differ in the central residue X, which is Aib, L-Ala, D-Ala, or Gly.⁵⁸ The presence of a central Aib residue clearly enhances hydrogen bonding, while a central Gly residue shows the least hydrogen bonding, as expected.

4.3. Conformational Cluster Analysis to Detect Structural Differences. The four conformational ensembles for the four seven-residue peptides were compared by performing a conformational cluster analysis for three pairs of trajectories; see Figure 3. It shows that the four ensembles are quite different. The Aib peptide shows a stronger propensity towards bent structures, while the L-Ala peptide shows a tendency to adopt more extended conformations, and the Gly peptide shows preference for a β -turn.⁵⁸

4.4. Configurational Entropy as Function of Time. In Figure 4 the Schlitter entropy is shown as function of time for two differently protonated α - and β -peptides of similar lengths consisting purely of Ala amino acid residues and solvated in water and in methanol. On a time scale of 10 ns, the Schlitter entropy of the α -helical conformation of the eight-residue α -peptide and of the 3_{14} -helical conformation of the six-residue β -peptide is well-converged. Although all peptides contain the same number of 24 backbone atoms, the α -peptides show a significantly higher configurational entropy per atom than the β -peptides, irrespective of their protonation state or solvent.⁵⁹

5. CONCLUSION

An overview over the different types of analysis implemented in the GROMOS++ software has been given. Three types of analysis programs were distinguished: programs that calculate structural, dynamic, or thermodynamic quantities from configurational trajectories. Additional programs are available that compute ensemble averages of experimentally observable quantities or quantities derived from experimental data, e.g. NMR NOE intensities or atom–atom distance bounds, 3J -coupling constants, residual dipolar couplings (RDCs) and order parameters, small- and wide-angle X-ray scattering (SAXS/WAXS) intensities, and neutron scattering intensities for liquids.⁴²

Compared to self-written, individual scripts computing such quantities, the use of GROMOS++ offers the advantage that most of the building blocks required for a particular type of analysis are available and tested by a wide group of users. Composing new analysis programs is relatively straightforward because of the implementation of atom, vector, and property specifiers, which allow for a very flexible description of individual quantities to be calculated. Each GROMOS++ analysis program is described in the GROMOS manual and in digital, in-code documentation. Changing the code to add new

functionality does not need much effort, since GROMOS++ is written in C++ to support maximal reusability of source code.³⁹

In summary, GROMOS++ is a flexible and rich collection of analysis tools ready to be used for a variety of types of analysis regarding molecular simulation trajectories.

APPENDIX: THE GROMOS++ PROGRAMS

Table A1. List of GROMOS++ Programs for Preprocessing of a Molecular Simulation

name	description
bin_box	creates a configuration of a condensed phase system consisting of two components
build_box	generates a configuration of a condensed phase system on a grid (only one component)
check_box	checks the box dimensions of a trajectory file
check_top	checks a molecular topology for (consistency) errors
com_top	combines (multiple) molecular topology files into one
con_top	converts a molecular topology to one based on a different force-field version
copy_box	repeats/extends a simulation box along a given Cartesian axis
cry	performs (crystallographic) symmetry operations on configurations of molecules
explode	places molecules of a given box on a grid thereby expanding intermolecular distances to satisfy a specific minimum intermolecular distance
gca	generates Cartesian coordinates for atoms from specified distances and/or (dihedral) angles for the atoms
gch	generates Cartesian coordinates for hydrogen atoms based on the coordinates of covalently bound neighbor atoms
ion	replaces solvent molecules by ions based on the local electrostatic potential or by random selection
make_pt_top	takes two or more molecular topologies and writes the differences in the perturbation topology format
make_sasa_top	adds the SASA block to a molecular topology file
make_top	creates a molecular topology file
mk_script	generates the scripts and input files to run a molecular simulation
pdb2g96	converts coordinate files from pdb to GROMOS format
pert_top	creates a perturbation topology to uniformly set interactions to given values for specified atoms
prep_eds	generates dual molecular and perturbation topologies for an EDS simulation
pt_top	combines molecular topologies and perturbation topologies to write new (perturbation) topologies
ran_box	creates a configuration for a condensed phase system of any composition with random molecule placements
ran_solvation	solvates a solute by randomly placing solvent molecules around it
red_top	reduces a molecular topology to one for a subset of atoms
sim_box	solvates a solute in a solvent box removing solvent molecules that are too close to solute atoms

Table A2. List of GROMOS++ Programs for Postprocessing of a Molecular Simulation

name	description
bilayer_dist	computes the atom distribution along a bilayer normal to characterize a membrane system
bilayer_oparam	calculates order parameters for bilayer systems (membranes) with respect to a fixed orientation (usually the bilayer normal)
cluster	performs a conformational clustering based on a RMSD matrix, e.g., calculated by the program rmsdmat
cog	calculates the center of geometry or center of mass position of all solute atoms of a simulation trajectory
dfmult	calculates free energy differences between multiple states A from a simulation at a reference state R
dg_ener	calculates the free energy difference between two states A and B, based on the perturbation formula; reads the output of the program ener
diffus	calculates the diffusion constant for a selected set of atoms
dipole	calculates the electric dipole moment for a selected set of atoms
ditrans	monitors transitions of torsional dihedral angle rotations with respect to the potential energy
dssp	detects secondary structure elements in a protein according to the Kabsch and Sander rules ⁴⁹
eds_mult_all	calculates the parameters needed for an enveloping distribution sampling (EDS) simulation from energy time series, based on an iterative scheme
edyn	performs an essential dynamics analysis over a trajectory file; the covariance matrix is calculated and diagonalized for specified atoms
ene_ana	writes a time series for specific values from a (free) energy trajectory file; simple statistics or calculations of combined trajectory entries are possible
ener	recalculates user specified interaction energies from molecular trajectory files using the interaction parameters from the molecular topology
eps_field	calculates the relative dielectric permittivity from a trajectory of a molecular simulation in which an external electric field was applied
epsilon	calculates the relative dielectric permittivity based on a Kirkwood–Fröhlich type of equation (fluctuation formula)
filter	reduces/filters a coordinate trajectory to contain only a specified set of atoms
follow	creates a three-dimensional trace of selected atoms through time; the program takes the nearest image with respect to the previous atom position
gathtraj	gathers a trajectory using the specified gathering method
hbond	monitors the occurrence of two- and three-centered hydrogen bonds
int_ener	recalculates the nonbonded interaction energy between two nonoverlapping sets of solute atoms using the interaction parameters specified in the molecular topology
iondens	calculates the average density of ions (or other particles) in space from a molecular trajectory file
jepot	computes the ³ J-coupling local elevation (LE) potential from a LE ³ J-coupling restrained simulation
jval	generates time series of ³ J-coupling constants based on a molecular trajectory
m_widom	calculates the free energy of inserting a test particle into configurations of a molecular system
matrix_overlap	calculates the overlap of two matrices (a mathematical definition of the overlap is given in the digital in-code documentation of GROMOS++)
mdf	for a given central set of atoms, mdf calculates the distance to the nearest atom belonging to a second set of atoms
nhoparam	calculates NH-order parameters from a simulation trajectory
noe	calculates and averages atom–atom distances; the trajectories originate either from a NOE distance restrained or free molecular simulation; the analysis may need preprocessing of data using the program prep_noe
post_noe	reanalysis of data generated by the program noe, resulting in NOE-bound violations
postcluster	performs lifetime-analysis, combined clustering, and writing of coordinates of (central) members of clusters, based on the output of the program cluster
prep_noe	converts X-plor NOE data formats to the GROMOS++ format (preparation for the noe program)
rdf	calculates a radial distribution function for specified atoms
rep_ana	used for analysis of molecular replica exchange simulations
rep_rewrite	sorts replica exchanged trajectories according to the λ or temperature values and writes them to different sorted files
reweight	reweights a time series of observed values of a quantity X sampled during a simulation at state R, i.e., using the Hamiltonian $H_R(\mathbf{p}, \mathbf{r})$, to another state Y (neglecting kinetic contributions for simplicity)
rgyr	calculates the radius of gyration for a specified set of atoms
rmsd	calculates the atom-positional root-mean-square deviation of a selected set of atoms
rmsdmat	calculates the positional root-mean-square deviation matrix for a set of structures; the output may be analyzed by the program cluster
rmsf	computes the positional root-mean-square fluctuations for a specified set of atoms
sasa	calculates the solvent-accessible surface area for selected atoms using the algorithm described by Lee and Richards ⁴⁷
sasa_hasel	calculates the solvent-accessible surface area using Hasel's formula ⁴⁸
solute_entropy	calculates the configurational entropy based on a coordinate trajectory
tcf	calculates distributions and time correlation functions
tser	calculates time series of quantities
tstrip	removes solvent coordinates from a simulation trajectory
visco	calculates the bulk and shear viscosities

AUTHOR INFORMATION

Corresponding Author

*E-mail: wfvgn@igc.phys.chem.ethz.ch.

ACKNOWLEDGMENT

Ulf Börjesson, Roland Bürgi, Markus Christen, Alice Glättli, Mika Kastenholz, Christine Peter, Heiko Schäfer, Daniel Trzesniak and Haibo Yu are kindly acknowledged for their previous work on GROMOS++, which would not be at its current state without their efforts. This work was financially supported by the National Center of Competence in Research (NCCR) in Structural Biology, by grant number 200020-121913 of the Swiss National Science Foundation, and by grant number 228076 of the European Research Council, which are gratefully acknowledged.

REFERENCES

- (1) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064.
- (2) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57.
- (3) van Gunsteren, W. F.; Dolenc, J. Biomolecular simulation: Historical picture and future perspectives. *Biochem. Soc. Trans.* **2008**, *36*, 11.
- (4) Lindahl, E.; Sansom, M. S. P. Membrane proteins: Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 425.
- (5) Khalili-Araghi, F.; Gumbart, J.; Wen, P.-C.; Sotomayor, M.; Tajkhorshid, E.; Schulten, K. Molecular dynamics simulations of membrane channels and transporters. *Curr. Opin. Struct. Biol.* **2009**, *19*, 128.
- (6) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The AMBER biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668.
- (7) Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545.
- (8) Bowers, K. J.; Chow, E.; Huageng Xu; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, FL, 2006.
- (9) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (10) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719.
- (11) Banks, J. L.; Beard, H. S.; Cao, Y. X.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* **2005**, *26*, 1752.
- (12) Lee, F. S.; Chu, Z. T.; Warshel, A. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYME programs. *J. Comput. Chem.* **1993**, *14*, 161.
- (13) Chu, Z. T.; Villa, J.; Strajbl, M.; Schutz, C. N.; Shurki, A.; Washel, A. MOLARIS v. beta9.05; University of Southern California, 2002.
- (14) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781.
- (15) TINKER—Software tools for molecular design, <http://dasher.wustl.edu/tinker> (visited on August 4, 2011).
- (16) Ponder, J. W.; Case, D. A. Force fields for protein simulations. In *Protein Simulations*; Academic Press Inc.: San Diego, CA, 2003; pp 27.
- (17) Cheatham, T. E.; Young, M. A. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers* **2000**, *56*, 232.
- (18) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586.
- (19) MacKerell, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400.
- (20) Chen, J. H.; Im, W. P.; Brooks, C. L. Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field. *J. Am. Chem. Soc.* **2006**, *128*, 3728.
- (21) MacKerell, A. D.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257.
- (22) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671.
- (23) Patel, S.; MacKerell, A. D.; Brooks, C. L. CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* **2004**, *25*, 1504.
- (24) Lamoureux, G.; Roux, B. Modeling induced polarization with classical drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.* **2003**, *119*, 3025.
- (25) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **2006**, *418*, 245.
- (26) Arnautova, Y. A.; Jagielska, A.; Scheraga, H. A. A new force field (ECEPP-05) for peptides, proteins, and organic molecules. *J. Phys. Chem. B* **2006**, *110*, 5025.
- (27) Levitt, M.; Hirshberg, M.; Sharon, R.; Daggett, V. Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. *Comput. Phys. Commun.* **1995**, *91*, 215.
- (28) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solution*; John Wiley & Sons, Inc.: New York, 1991.
- (29) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205.
- (30) Chandrasekhar, I.; Kastenholz, M.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; Tieleman, D. P.; van Gunsteren, W. F. A consistent potential energy parameter set for lipids: Dipalmitoylphosphatidylcholine as a benchmark of the GROMOS96 45A3 force field. *Eur. Biophys. J. Biophys.* **2003**, *32*, 67.
- (31) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Kräutler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F. An

improved nucleic acid parameter set for the GROMOS force field. *J. Comput. Chem.* **2005**, *26*, 725.

(32) Lins, R. D.; Hünenberger, P. H. A new GROMOS force field for hexopyranose-based carbohydrates. *J. Comput. Chem.* **2005**, *26*, 1400.

(33) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656.

(34) Poger, D.; van Gunsteren, W. F.; Mark, A. E. A new force field for simulating phosphatidylcholine bilayers. *J. Comput. Chem.* **2010**, *31*, 1117.

(35) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. Definition and testing of the GROMOS force-field versions: 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843.

(36) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(37) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474.

(38) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. The GROMOS biomolecular simulation package. *J. Phys. Chem. A* **1999**, *103*, 3596.

(39) Schmid, N.; Christ, C. D.; Christen, M.; Eichenberger, A. P.; van Gunsteren, W. F. Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. *Comput. Phys. Commun.* **2011** submitted.

(40) Kunz, A. P. E.; Allison, J. R.; Geerke, D. P.; Horta, B. A. C.; Hünenberger, P. H.; Riniker, S.; Schmid, N.; van Gunsteren, W. F. New functionalities in the GROMOS biomolecular simulation software. *J. Comput. Chem.* **2011** submitted.

(41) Riniker, S.; Christ, C. D.; Hansen, H.; Hünenberger, P. H.; Oostenbrink, C.; Steiner, D.; van Gunsteren, W. F. Computation of relative free energies for ligand protein binding, solvation and conformational transitions using the GROMOS biomolecular simulation software. *J. Phys. Chem.* **2011**, submitted.

(42) Schmid, N.; Allison, J. R.; Dolenc, J.; Eichenberger, A. P.; Kunz, A. P. E.; van Gunsteren, W. F. Biomolecular structure refinement using the GROMOS simulation software. *J. Biomol. NMR* **2011**; DOI: 10.1007/s10858-011-9534-0.

(43) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-integration of Cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327.

(44) Allison, J. R.; Boguslawski, K.; Fraternali, F.; van Gunsteren, W. F. A refined, efficient mean solvation force model that includes the interior volume contribution. *J. Phys. Chem. B* **2011**, *115*, 4547.

(45) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300.

(46) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.: Zürich, Groningen, 1996.

(47) Lee, B.; Richards, F. M. Interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379.

(48) Hasel, W.; Hendrickson, T. F.; Still, W. C. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput. Methodol.* **1988**, *1*, 103.

(49) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577.

(50) Koehler, J. E. H.; Saenger, W.; van Gunsteren, W. F. On the occurrence of three-center hydrogen bonds in cyclodextrins in crystalline form and in aqueous solution: Comparison of neutron diffraction and molecular dynamics results. *J. Biomol. Struct. Dyn.* **1988**, *6*, 181.

(51) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding-unfolding thermodynamics of a β -heptapeptide from equilibrium simulations. *Proteins* **1999**, *34*, 269.

(52) Riniker, S.; Kunz, A. P. E.; Gunsteren, W. F. On the calculation of the dielectric permittivity and relaxation of molecular models in the liquid phase. *J. Chem. Theory Comput.* **2011**, *7*, 1469.

(53) Widom, B. Some topics in theory of fluids. *J. Chem. Phys.* **1963**, *39*, 2808.

(54) Liu, H. Y.; Mark, A. E.; van Gunsteren, W. F. Estimating the relative free energy of different molecular states with respect to a single reference state. *J. Phys. Chem.* **1996**, *100*, 9485.

(55) Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. nonpolar gases. *J. Chem. Phys.* **1954**, *22*, 1420.

(56) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187.

(57) Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance-matrix. *Chem. Phys. Lett.* **1993**, *215*, 617.

(58) Wang, D.; Friedmann, M.; Gattin, Z.; Jaun, B.; van Gunsteren, W. F. The propensity of α -aminoisobutyric acid (=2-methylalanine; Aib) to induce helical secondary structure in an α -heptapeptide: A computational study. *Helv. Chim. Acta* **2010**, *93*, 1513.

(59) Allison, J. R.; Müller, M.; van Gunsteren, W. F. A comparison of the different helices adopted by α - and β -peptides suggests different reasons for their stability. *Protein Sci.* **2010**, *19*, 2186.

On the Role of Water Models in Quantifying the Binding Free Energy of Highly Conserved Water Molecules in Proteins: The Case of Concanavalin A

Elisa Fadda[†] and Robert J. Woods^{*,†,‡}

[†]School of Chemistry, National University of Ireland Galway, University Rd, Galway, Ireland

[‡]University of Georgia, Complex Carbohydrate Research Center, 315 Riverbend Rd, Athens, Georgia 30602, United States

 Supporting Information

ABSTRACT: The ability of ligands to displace conserved water molecules in protein binding sites is of significant interest in drug design and is particularly pertinent in the case of glycomimetic drugs. This concept was explored in previous work [Clarke et al. *J. Am. Chem. Soc.* **2001**, *123*, 12238–12247 and Kadirvelraj et al. *J. Am. Chem. Soc.* **2008**, *130*, 16933–16942] for a highly conserved water molecule located in the binding site of the prototypic carbohydrate-binding protein Concanavalin A (Con A). A synthetic ligand was designed with the aim of displacing such water. While the synthetic ligand bound to Con A in an analogous manner to that of the natural ligand, crystallographic analysis demonstrated that it did not displace the conserved water. In order to quantify the affinity of this particular water for the Con A surface, we report here the calculated standard binding free energy for this water in both ligand-bound and free Con A, employing three popular water models: TIP3P, TIP4P, and TIP5P. Although each model was developed to perform well in simulations of bulk-phase water, the computed binding energies for the isolated water molecule displayed a high sensitivity to the model. Both molecular dynamics simulation and free energy results indicate that the choice of water model may greatly influence the characterization of surface water molecules as conserved (TIP5P) or not (TIP3P) in protein binding sites, an observation of considerable significance to rational drug design. Structural and theoretical aspects at the basis of the different behaviors are identified and discussed.

INTRODUCTION

Protein–carbohydrate recognition is a fundamental step in numerous essential biological processes, such as cell–cell and cell–matrix interactions, and conversely can be exploited in viral and bacterial adhesion and infection.¹ Detailed knowledge of the precise relationships between glycan structure and protein recognition requires the ability to relate atomic structures to thermodynamic properties. Such insight is paramount in the rational design of therapeutic agents that specifically target these interactions. The earliest carbohydrate-binding proteins to have received extensive analyses are the plant lectins.^{2,3} Because of the large amount of structural, thermodynamic, and mutational data that have been accumulated for lectins, they serve as excellent models for probing the origins of the specificity of carbohydrate–protein binding.^{1,3} A common feature in carbohydrate–protein complexes is the presence of discrete water molecules that appear to mediate the interaction through hydrogen bonding.² Such water molecules can also be found to occupy the same hydration sites in both the free and bound forms of the protein, suggesting that they play functional roles in recognition and binding. The presence of highly conserved water molecules in protein binding sites is of significant interest in drug design because there is the potential for a therapeutic agent to gain binding free energy through entropic gains relative to the native ligand upon their displacement.^{4–10} The rational design of ligands that can displace waters from specific protein binding sites requires a knowledge of accurate values of the binding free energies of such water molecules; this information can only be obtained via computer

simulation techniques. Despite their apparent significance, the affinities of these conserved waters in carbohydrate–protein complexes remain unknown.

The ability to define a relationship between binding free energies of conserved water molecules and ligand affinity is also of great interest. In their work, Barillari et al.¹¹ analyzed the binding free energies of several conserved and displaced water molecules in different proteins complexed with a variety of ligands. Nonetheless, they were not able to establish any direct correlation between water binding free energies and ligand affinity. If such a correlation exists, it may well not be generalizable to all proteins and all ligands. It is likely that the gain or loss of binding affinity due to water displacement is system-specific, as well as the balance between entropic and enthalpic binding contributions for the conserved water molecules.^{8,9} It is also important to recall that some water molecules are highly conserved because they play a structural role in the scaffold to which the ligand binds. Therefore, as in the mutation of key protein residues in the binding site, it is foreseeable that the displacement of these water molecules could also destabilize the ligand.

Con A, isolated from the jack bean (*Canavalia ensiformis*), is not only one of the best-characterized lectins but also shows a well-defined example of water-mediated carbohydrate binding. Isothermal titration microcalorimetry (ITC) experiments have

Received: June 14, 2011

Published: August 23, 2011

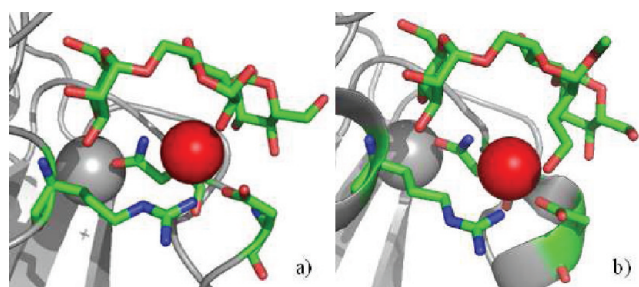


Figure 1. (a) Detail of the saccharide binding site of Concanavalin A (Con A) bound to 3MAN (PDBID 1CVN, ref 14). (b) Con A bound to 3HET (PDBID 3D4K, ref 25). Structural water is shown as a red sphere, while ligands and key residues, i.e., ASN 14, ASP 16, and ARG 228, are shown as sticks. Oxygen atoms are shown in red, carbon in green, and nitrogen in blue. The remainder of the protein is shown in gray, with the structural metal ions shown as spheres.

established that Con A binds the trisaccharide α -D-Manp(1–6)-[α -D-Manp(1–3)] α -D-Manp (3MAN), present in all *N*-linked glycans, with a 60-fold higher affinity than the monomer methyl- α -D-mannopyranose.¹² The high affinity of Con A for the trimannoside core has been the subject of numerous studies.^{12–20} X-ray crystallographic data reveal the presence of specificity-defining hydrogen bonds, as well as nonspecific hydrophobic interactions, which involve all three carbohydrate residues and extend along the length of the binding site.¹⁴ The conserved water participates directly in the hydrogen bond network (Figure 1). This water molecule is tetrahedrally coordinated to the O2 of the central (reducing) sugar in 3MAN, and to the protein residues ASN 14, ASP 16, and ARG 228.¹⁴ A water molecule in the same location is also observed in crystal structures of the unbound protein.^{21–23} Earlier thermodynamic,²⁴ crystallographic,²⁵ and theoretical studies^{17,24–26} explored the possibility of displacing this water by means of a synthetic analog of 3MAN, namely 3HET (Figure 1, panel b). In 3HET, a hydroxyethyl group replaces the hydroxyl group at the C2 position of the central mannosyl residue. ITC data²⁴ showed a more favorable entropy contribution for the binding of 3HET to Con A, relative to 3MAN, which was initially interpreted as indicating the displacement of the conserved water. A subsequent crystallographic and computational study²⁵ showed that the water was not displaced, although its position was slightly distorted relative to that in the Con A–3MAN complex (Figure 1). In the Con A–3HET complex, the conserved water forms a similar hydrogen bond network as present in the Con A–3MAN complex. In order to determine the molecular basis for the inability of the synthetic ligand to displace the water and to determine whether that particular water could reasonably be targeted for displacement in the first place, we have computed binding free energies for the water in the unbound protein, as well as in the protein bound to the natural and synthetic ligands. Additionally, extended (100 ns) unrestrained molecular dynamics (MD) simulations were run on all systems to investigate this water's residence times, and water–protein–ligand dynamics.

The interpretation of subtle changes in predicted free energy values places strenuous demands on the precision of the calculations. In this respect, equilibrium free energy calculation methods, such as thermodynamic integration (TI) used in this work, in combination with extensive sampling, are among the most reliable. In order to gauge the influence of different water models

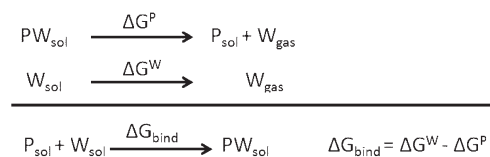


Figure 2. Thermodynamic cycle used to derive standard binding free energies for the conserved water in the Con A saccharide binding site. P_{sol} indicates the protein (Con A) in solution without the conserved water molecule, and $P_{\text{sol}} + W_{\text{sol}}$ indicates Con A with the bound water molecule in solution. W_{sol} indicates a water molecule in bulk water, and W_{gas} indicates a water molecule in the gas phase. ΔG^P is the free energy required for removing a water from the protein binding site, and ΔG^W is the hydration free energy.

on the binding free energy calculations, we focused on three popular nonpolarizable models: TIP3P,²⁷ TIP4P,²⁷ and TIP5P,²⁸ which employ three, four, or five sites to describe the electrostatic properties of water, respectively.

Our results demonstrate an unexpected high sensitivity to the water model used in free energy calculations, with the TIP5P water model being the only one that characterizes the water as bound both in the free protein and in all its complexes. Additional unrestrained MD simulations (100 ns) show that in the free protein TIP3P and TIP4P waters exchange with bulk water far more rapidly than TIP5P. Similar conclusions have been recently reported based on peptides simulations.²⁹ In addition, the TIP3P and TIP4P waters continue to exchange, albeit more slowly, even in the presence of the bound ligands, while for the duration of the simulations the same TIP5P water remains bound in both complexes. Notably, and in contradiction with the structural data,²⁵ the MD simulations with TIP3P and TIP4P show that the hydroxyethyl moiety in the 3HET ligand spontaneously displaces both bound waters. In contrast, the TIP5P water resists displacement by 3HET.

MATERIALS AND METHODS

Standard binding free energies were determined by double decoupling (DD),³⁰ based on the diagram shown in Figure 2. In the DD method, the free energy for ligand binding $\Delta G_{\text{bind}}^{\circ}$ is computed as the difference between the hydration free energies for the ligand in bulk water ΔG_{TI}^W and the decoupling free energy of the ligand in the protein ΔG_{TI}^P , where the term “decoupling” refers to the stepwise disappearance of the ligand, computed here using thermodynamic integration (TI):

$$\Delta G_{\text{bind}}^{\circ} = \Delta G_{\text{TI}}^W (\Delta G_{\text{TI}}^P + \Delta G_{\text{Corr}}) \quad (1)$$

ΔG_{Corr} is the standard state correction term required to correct for a standard concentration (C°) of 1 M, which in molecular simulations is more conveniently expressed as 1 molecule/1660 \AA^3 . In order to prevent the diffusion of the water out of the binding site during its annihilation, the oxygen of the water molecule is harmonically restrained throughout the calculations. The choice of the restraining potential follows the directives suggested in work by Hamelberg and McCammon.¹⁰ This restraint restricts the volume sampled by the water, and it is accounted for in the correction term:³⁰

$$\Delta G_{\text{Corr}} = -RT \frac{\sigma_{\text{PW}}}{\sigma_{\text{P}} \sigma_{\text{W}}} + RT [C^{\circ} (2\pi RT/k)^{3/2}] \quad (2)$$

where σ_{PW} , σ_{P} , and σ_{W} are the symmetry numbers for the protein–water complex (PW), for the protein (P), and for the

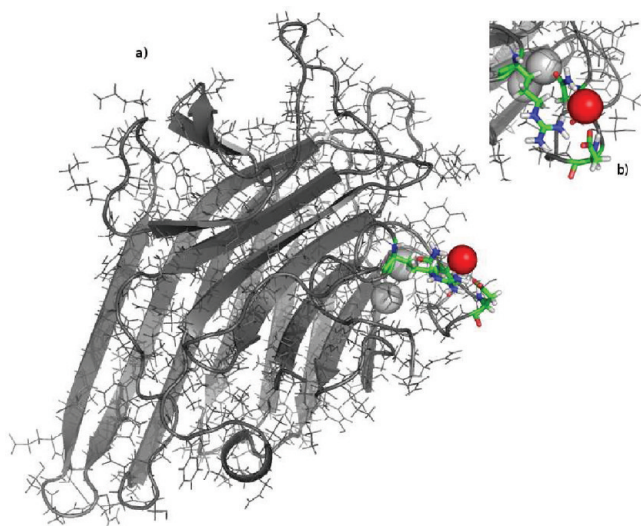


Figure 3. (a) Con A (monomer; PDBid 1GKB, ref 40). (b) Detail of the saccharide binding site with the key residues, ASN 14, ASP 16, and ARG 228 highlighted. The structurally conserved water is represented by a red van der Waals sphere, the metal ions by two gray spheres.

water molecule (W), respectively, and k is the harmonic force constant ($k = 2.4 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). The standard state correction term ΔG_{corr} in our calculations amounts to $-3.0 \text{ kcal mol}^{-1}$. The same harmonic restraints were applied in all DD calculations.

The following crystal structures, indicated here by their PDBid, were chosen as starting conformations: 1CVN¹⁴ for Con A bound to the trisaccharide 3MAN, 3D4K²⁵ for Con A bound to the synthetic analog 3HET, and 1GKB²¹ for free Con A. At pH = 7.0, Con A exists both as a dimer and as a tetramer.³¹ However, the ligand binding sites are independent of multimeric structure, and only one of the monomers was used in all simulations (see Figure 3). Two metal ions, a Mn(II) and a Ca(II), are required for Con A's structural stability and for ligand binding. Due to the lack of force field parameters for Mn(II), both metals were described as Ca(II). In addition to TIP3P, TIP4P, and TIP5P, the performance of a rarely used modified version of the TIP3P model, namely TIP3P-MOD,³² was also assessed. The latter differs from TIP3P exclusively in the values of the Lennard-Jones (LJ) parameters. The results obtained are very similar to the ones obtained with the TIP3P model and are included as Supporting Information.

The following simulation protocol was followed in all cases. Crystal structures of the protein were inserted in an equilibrated cubic water box (80 Å cell length). The *genion* routine, included in the GROMACS software package, was used to add the three Na⁺ counterions necessary to neutralize the system. The positions of the added water molecules, counterions, and hydrogen atoms were minimized with the steepest descent algorithm in 15 000 steps. An initial 500 ps equilibration at constant volume and temperature (the NVT at 300 K) was performed for the added water molecules, counterions, and hydrogen atoms. This step was followed by another 500 ps equilibration step at atmospheric constant pressure (NPT at 300 K). Protein, ligand, and conserved water were then equilibrated for 1 ns in the NPT ensemble, with the position of the protein C-alpha (C α) atoms, heavy atoms of the ligand, and the oxygen (OW) of the conserved water restrained

harmonically with a force constant of $2.4 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$. This equilibration step was followed by a final production step of 10 ns. The conformations obtained after the 10 ns MD simulation were used as the starting point for the free energy calculations. Thermodynamic integration (TI) in bulk water was performed with 20 lambda (λ) points, while decoupling in the protein binding site involved integration over 37 λ points. These steps were separated by 0.01 between 0.00 and 0.10 and between 0.90 and 1.00 and by 0.05 between 0.10 and 0.90. A finer separation around $\lambda = 0$ and $\lambda = 1$ was necessary to reduce the noise experienced at the beginning and end of the decoupling process and therefore to improve precision. In all cases, the Coulomb interactions were decoupled first, followed by the vdW interactions. At each λ value, an energy minimization was performed for 15 000 steps of steepest descent, followed by equilibration for 1 ns (NPT). Production was run for 1 ns. Tests with longer production runs, i.e., 5 ns per λ and even 10 ns per λ , did not show significant differences in the results. The final free energy values are based on a total simulation time of 72 ns for each system. During TI, harmonic restraints were applied to the positions of all heavy atoms in the ligand; the OW of the water; and heavy atoms of residues ASN 14, ASP 16, and ARG 228. In particular, ASP 16 and ARG 228 are not only directly hydrogen-bonded to the conserved water but also form a salt bridge. Restraining the salt bridge in a closed position, which corresponds to the crystallographic conformation in all structures available in the literature, was found to be necessary because its opening and closing greatly affects free energy values. Water molecules from the bulk can access the protein binding site as the conserved water is completely decoupled. In similar studies, a square-well potential was applied in order to prevent this from happening.^{5,11} When Con A is bound to a ligand, either 3MAN or 3HET, the binding site is not accessible to bulk water, even when the conserved water molecule is annihilated. As the option of applying an external potential is not available in the software package we used (see below for details), in the case of free Con A, we confirmed that no bulk water diffused into the binding site during decoupling by analyzing the trajectories. In order to avoid end-point singularities, the vdW interactions were treated via soft-core interactions with a soft-core parameter $\alpha = 0.5$, an exponent of 1, and $\sigma = 0.3 \text{ nm}$. Integration was performed with the *scipy* module of python 2.6 (<http://www.python.org/>). For all MD simulations performed in the NPT ensemble, the temperature was held constant at 300 K by a Langevin thermostat with a coupling time constant of 0.1 ps. Pressure was held at 1 bar by exponential relaxation pressure coupling with a time constant of 0.5 ps. All MD simulations were performed with versions 4.0.3 up to 4.0.7 of the GROMACS software package.³³ The equations of motion were integrated using a leapfrog stochastic dynamics integrator with a 2 fs time step. Long-range electrostatics were treated with the particle mesh Ewald (PME) method.^{34,35} The maximum spacing for the fast fourier transform grid was chosen as 1 Å. Cutoff values for Coulomb and vdW interactions were set to 0.9 nm. Tests with longer cutoffs values (12 Å) were performed, but no significant differences were observed. Water hydration free energies were calculated in a cubic water box of 25 Å per side. The calculated water densities at 300 K are 0.985 g cm^{-3} for TIP3P, 0.993 g cm^{-3} for TIP4P, and 0.983 g cm^{-3} for TIP5P. A separate set of MD simulations with restraints only on the C α and on the Ca²⁺ ions bound to the protein were performed for

100 ns for all systems with all three water models for a total of 900 ns simulation time.

The AMBER99SB force field³⁶ was chosen to represent the protein, while both carbohydrate ligands were represented with the 06g release of the GLYCAM06 force field.³⁷

RESULTS AND DISCUSSION

1. Water Hydration Free Energies. The hydration free energies for each water model are shown in Table 1. The TIP5P hydration free energy is slightly lower compared not only to TIP3P and TIP4P but also to other popular water models.³⁸ The values reported in the literature range between an upper bound of -7.0 kcal mol⁻¹ for the SPC/E model to a lower bound of -6.1 kcal/mol^{38,39} for TIP3P and TIP4P. The agreement between the TIP3P and TIP4P hydration free energy values here and the values reported in the literature confirm the accuracy of the TI protocol employed in this study. Variations in the electrostatics contribute most significantly to the differences in the hydration free energies between TIP3P, TIP4P, and TIP5P. As shown in Table 1, the free energy required to annihilate the electrostatic interactions is the same for TIP3P and TIP4P, i.e., 8.5 and 8.4 kcal mol⁻¹, respectively. For TIP5P, the Coulomb contribution is weaker, i.e., 7.7 kcal mol⁻¹. The van der Waals (vdW) contributions are similar and unfavorable for all water models, ranging from 2.0 to 2.3 kcal/mol. These contributions reflect the steric compression resulting from the strongly attractive electrostatic interactions in liquid water.

Electrostatic and Lennard-Jones parameters characterizing TIP3P, TIP4P, and TIP5P are shown in Table S.1 (Supporting Information).

2. Water Binding Free Energy in Free Con A. In free Con A, the conserved water is coordinated to the binding site's key residues, ASN 14, ASP 16, and ARG 228, through three hydrogen bonds. The fourth hydrogen bond is directed to a water molecule in the bulk (see Figures 3 and 4).⁴⁰ ASP 16 and ARG 228 form a salt bridge. Table S.2 (Supporting Information) shows the hydrogen bond distances in the crystal structure (1GKB) and in the free Con A snapshots used to start the free energy calculations. These snapshots were obtained from 10 ns MD

simulations, where the oxygen of the conserved water was harmonically restrained to its crystallographic position. Such restraint was necessary and sufficient to ensure that water's coordination was preserved, as well as the closed conformation of the salt bridge between ASP 16 and ARG 228. The coordination of TIP3P, TIP4P, and TIP5P in free Con A is shown in Figure 4. All water models are able to reproduce correctly the water coordination, with the only notable difference being that the salt bridge between the carboxylate oxygen in ASP 16 and the amino nitrogen in ARG 228 is shortened in the MD snapshots, and especially in the structures with TIP3P and TIP5P water, relative to the crystal structure.

The binding free energies for the conserved water in free Con A are shown in Table 2. The free energy cost for displacing the conserved water molecule from Con A's binding site is -2.3 kcal mol⁻¹ for both TIP5P and TIP4P. This value is consistent with crystallographic evidence^{22,23,40} indicating the presence of a water molecule in the binding site of the free protein. Conversely, the value of $+0.1$ kcal mol⁻¹ obtained for TIP3P does not suggest occupancy of the binding site. The modification of the TIP3P Lennard-Jones parameters characterizing the TIP3P-MOD model does not correct the TIP3P behavior, i.e., -0.3 kcal mol⁻¹ (see Table S.4 in Supporting Information).

3. Water Binding Free Energy in the Con A–3MAN Complex. As in free Con A, the conserved water molecule forms four hydrogen bonds also in the binding site of the Con A–3MAN complex. The crystal structure 1CVN indicates that three hydrogen bonds connect the water to the key residues, ASN 14, ASP 16, and ARG 228, while the fourth hydrogen bond is directed to the O2 of the central (reducing) mannosyl residue of the ligand. This water coordination is generally well preserved in all of the structures used as a starting point for the free energy calculations, with only small differences in the hydrogen bonding coordination pattern (see Figure 5). Specifically, while one lone pair in TIP5P forms a hydrogen bond with the ARG 228 imino group ($-NH-$), both TIP3P and TIP4P interact with ARG 228 through one of the two amino group protons ($-NH_2$; see Figure 5 and Table S.2, Supporting Information, for hydrogen bond distances). The presence of the two lone pairs in TIP5P clearly facilitates the formation of linear hydrogen bonds with

Table 1. Hydration Free Energies^a (ΔG°) and Individual Contributions Obtained for the TIP3P, TIP4P, and TIP5P Water Models

water model	Coulomb	vdW	ΔG°	literature
TIP3P	-8.5	2.2	-6.3 (0.1) ^b	-6.5, ³⁹ -6.1 ³⁸
TIP4P	-8.4	2.3	-6.1 (0.2)	-6.1 ³⁸
TIP5P	-7.7	2.0	-5.7 (0.1)	

^a All energies are in kcal mol⁻¹. ^b Errors, calculated by block averaging, are shown in parentheses.

Table 2. Standard Binding Free Energies^a (ΔG°) and Individual Contributions for the Conserved Water in the Binding Site of Free Con A

free Con A	Coulomb	vdW	ΔG_p	ΔG°
TIP3P	-14.9	5.7	-6.2	0.1 (0.1) ^b
TIP4P	-13.7	2.3	-8.4	-2.3 (0.1)
TIP5P	-15.5	4.5	-8.0	-2.3 (0.2)

^a All energies are in kcal mol⁻¹. ^b Errors, calculated by block averaging, are shown in parentheses. ΔG_p includes a standard state correction term of 3.0 kcal mol⁻¹.

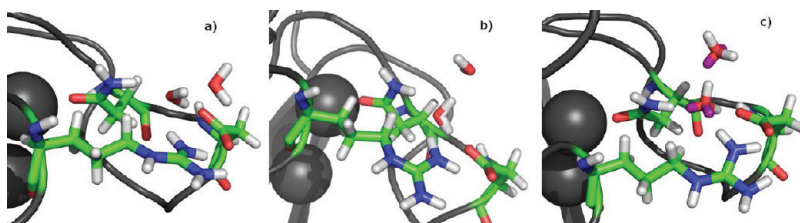


Figure 4. Coordination of TIP3P, TIP4P, and TIP5P in free Con A

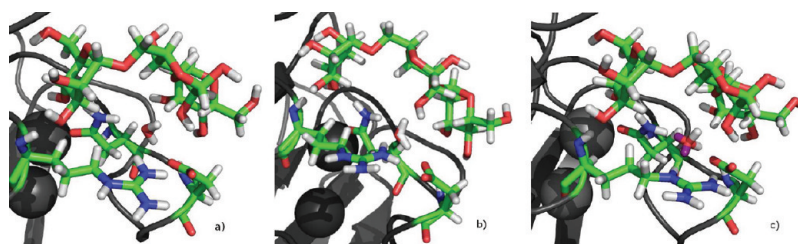


Figure 5. (a) Detail of the binding site in free Con A from the MD simulation with a TIP3P water representing the structurally conserved water. (b) Detail of the binding site in free Con A from the MD simulation with a TIP5P water representing the structurally conserved water. The TIP5P lone pairs are highlighted in magenta.

Table 3. Standard Binding Free Energies^a (ΔG°) and Individual Contributions for the Conserved Water in the Binding Site of Con A in Complex with 3MAN

Con A–3MAN	Coulomb	vdW	ΔG_p	ΔG°
TIP3P	−21.7	11.4	−7.3	−1.0 (0.2) ^b
TIP4P	−21.5	10.1	−8.4	−2.3 (0.3)
TIP5P	−21.1	5.3	−12.8	−7.1 (0.1)

^aAll energies are in kcal mol^{−1}. ^bErrors, calculated by block averaging, are shown in parentheses. ΔG_p includes a standard state correction term of 3.0 kcal mol^{−1}.

all residues. Conversely, the hydrogen bond between the TIP3P and TIP4P oxygen and the $-\text{NH}_2$ of ARG 228 is not as linear.

As shown in Table 3, in the Con A–3MAN complex, the TIP5P water has a much higher binding free energy value of -7.1 kcal/mol relative to free Con A. High standard binding free energy values for conserved water molecules are not unusual,¹¹ and they suggest a structural role for water molecules in ligand binding. The values obtained for TIP3P and TIP4P are much lower relative to TIP5P, i.e. -1.0 kcal/mol and -2.3 kcal/mol, respectively. As shown in Table 3, the steric compression, accounted for by the vdW decoupling free energy contributions, is much higher for TIP3P and TIP4P waters relative to TIP5P. Interestingly, the binding free energy value obtained for TIP4P is equal to the value calculated for the TIP4P water in the free protein, suggesting that interchanging bulk water molecules are as effective in coordinating the conserved water as is the O2 in 3MAN, that is, suggesting that the ligand does not enhance the stability of the bound water.

4. Water Binding Free Energy in the Con A–3HET Complex.

Crystallographic data show that the position of the water in the Con A–3HET complex is only slightly distorted relative to the one in the Con A–3MAN complex.²⁵ In the crystal structure of the Con A–3HET complex (3D4K), the conserved water is hydrogen-bonded to the same protein residues as it is in the Con A–3MAN complex, but it interacts with the O8 of the hydroxyethyl group instead of the O2 of the central mannosyl residue of 3MAN (see Figures 1 and 6). The conserved water coordination is reproduced correctly by all water models with a few slight differences. As shown in Figure 6 a and b, the coordinations of TIP3P and TIP4P are very similar with direct hydrogen bonds with ASN 14, ASP 16, ARG 228, and the O8 of the 3HET hydroxyethyl. In all structures of the Con A–3HET complex including the crystal structure, the hydrogen bond to the O8 of 3HET is slightly elongated relative to the hydrogen bond to O2 of 3MAN, suggesting a weaker interaction between

the ligand and the binding site (see Table S.2, Supporting Information).

Binding free energies for the water in the Con A–3HET complex are shown in Table 4. The binding free energy for a TIP3P water in the Con A–3HET complex is -4.8 kcal/mol, making TIP3P more tightly bound when Con A is in complex with the synthetic ligand relative to the complex with the natural ligand. The binding free energy obtained for the TIP4P water is $+0.2$ kcal mol^{−1}, which indicates no propensity for the binding site to be hydrated. The binding free energy of a TIP5P water in the Con A–3HET complex is 1.4 kcal mol^{−1} lower than the binding free energy in the Con A–3MAN complex, i.e., -5.7 kcal mol^{−1}. This value indicates that the water is still firmly bound within the binding site but in a less favorable location relative to the Con A–3MAN complex. Indeed, while all water models make stronger electrostatic contacts, in the Con A–3MAN complex, the steric compression of TIP3P and TIP4P is much higher than for TIP5P (see Tables 3 and 4).

5. Molecular Dynamics Simulations. The stability of the conserved water molecule and of the hydrogen bond network connecting it to the ligands and to the protein binding site was monitored for each water model throughout 100 ns of unrestrained MD simulations.

As seen in Figure 4, in the free protein, the conserved water is hydrogen-bonded to three key residues (ASN 14, ASP 16, and ARG 228) and to a water molecule from the bulk. The average hydrogen bond distances calculated over the 100 ns trajectories are shown in Table 5. The dynamics of the residues in the binding site and especially of the salt bridge between ASP 16 and ARG 228 promote the exchange between water molecules from the bulk and the water occupying the binding site. This exchange always occurs between the conserved water and the water hydrogen-bonded to it in a concerted manner so that the binding site never remains vacant. The exchange rate of the water molecules depends both on the water coordination and on the water model chosen. More specifically, as the salt bridge opens, the water loses one hydrogen bond either to ASP 16 or to ARG 228, and it is much more likely to exit the binding site. Most water exchanges are triggered by the opening of the salt bridge. As a complete kinetic analysis of the water exchange rates is beyond the scope of this work, we determined water molecule residence times on the basis of 100 ns of the lifetime of the hydrogen bond network between a specific water and the three key residues, ASN 14, ASP 16, and ARG 228. Residence times were calculated on the basis of 41 exchange events for TIP3P, 33 events for TIP4P, and 26 events for TIP5P. In order to calculate these values, the trajectories were extended from 100 ns up to 140 ns. In the binding site of free Con A, TIP5P has the longest residence time of 2.3 ns, while TIP3P and TIP4P residence times are 1.3 and 1.6 ns, respectively.

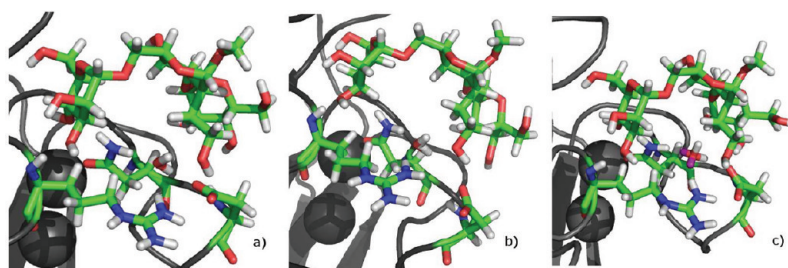


Figure 6. (a) Detail of the binding site in the Con A–3MAN complex from the MD simulation with a TIP3P water representing the structurally conserved water. (b) Detail of the binding site in the Con A–3MAN complex from the MD simulation with a TIP4P water representing the structurally conserved water. The TIP5P lone pairs are highlighted in magenta.

Table 4. Standard Binding Free Energies^a (ΔG°) and Individual Contributions for the Conserved Water in the Binding Site of Con A in Complex with 3HET

Con A–3HET	Coulomb	vdW	ΔG_p	ΔG°
TIP3P	−18.7	4.6	−11.1	−4.8 (0.1) ^b
TIP4P	−17.0	8.1	−5.9	+0.2 (0.4)
TIP5P	−19.0	4.6	−11.4	−5.7 (0.2)

^aAll energies are in kcal mol^{−1}. ^bErrors, calculated by block averaging, are shown in parentheses. ΔG_p includes a standard state correction term of 3.0 kcal mol^{−1}.

Table 5. Hydrogen Bond Distances^b between the Conserved Water and the Key Residues in the Binding Site of Free Con A^a

H-bond distances	1GKB	TIP3P	TIP4P	TIP5P
O _{WAT} –O _{ASP16}	2.6	2.7 (0.3) ^c	2.7 (0.2)	2.8 (0.3)
O _{WAT} –N _{ASN14}	2.9	3.1 (0.3)	3.0 (0.2)	2.9 (0.2)
O _{WAT} –N _{ARG228}	3.0	3.0 (0.2)	2.9 (0.2)	2.8 (0.1)

^aData from the crystal structure (1GKB) are compared to average distances calculated during the unrestrained MD simulations. ^bIn Ångstroms. ^cStandard deviations are indicated in parentheses.

The MD simulation of the Con A–3MAN complex shows that the presence of the trisaccharide limits the mobility of the water, reducing drastically the number of exchange events between the conserved water and water in the bulk. During the 100 ns simulations, a few exchange events are observed for both TIP3P and TIP4P water, but not for TIP5P, where the same water molecule occupied the binding site for the whole 100 ns. Average hydrogen bond distances are shown in Table 6. 3MAN is tightly bound to the protein surface throughout all simulations, and it shows a similar degree of mobility for all water models (RMSD values in Table S.5 and Figure S.1, Supporting Information). The terminal sugar residue more exposed to the bulk, namely, Man 3 (see Figure 7), shows a slightly higher degree of mobility relative to the others, especially in TIP4P. The slightly higher RMSD calculated for 3MAN in TIP4P is due to a partial detachment of the Man 3 residue at around 10 ns triggered by water exchange.

A partial detachment of the ligand from the protein was observed during the MD simulation of the Con A–3HET complex in TIP4P. As shown by the RMSD plot in Figure 8, between 15 and 21 ns, 3HET is highly mobile. During this time interval, 3HET remains bound to the protein only via Man 1. The partial detachment is caused by the conserved water

Table 6. Hydrogen Bond Distances^a between the Conserved Water and the Key Residues in the Binding Site of the Con A–3MAN Complex

H-bond distances	1CVN	TIP3P	TIP4P	TIP5P ^b
O _{WAT} –O _{ASP16}	2.8	2.6 (0.1) ^c	2.6 (0.1)	2.8 (0.2)
O _{WAT} –N _{ASN14}	2.7	2.9 (0.1)	3.0 (0.2)	2.8 (0.1)
O _{WAT} –N _{ARG228}	3.1	3.0 (0.2)	3.0 (0.2)	3.0 (0.3); 3.5(0.3)
O _{WAT} –O _{23MAN}	2.4	2.8 (0.2)	2.9 (0.2)	2.9 (0.2)

^aIn Ångstroms. ^bFor TIP5P, the first O_{WAT}–N_{ARG 228} value corresponds to the distance to the imino nitrogen of ARG 228, while the second one corresponds to the amino nitrogen. ^cStandard deviations are indicated in parentheses.

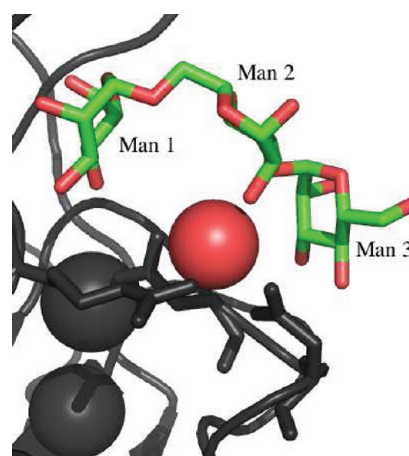


Figure 7. Residues nomenclature used for 3MAN. The conserved water is shown in red. The structure shown has PDBid 1CVN.¹⁴

leaving the binding site and not being immediately replaced. As a new water molecules enters the binding site at 21 ns, the 3HET returns to its bound conformation, where it stays until the end of the simulation. RMSD values for the ligand calculated over the 100 ns simulation are shown in Table S.6 (Supporting Information).

An interesting result obtained from the MD simulations on the Con A–3HET complex is that the hydroxyethyl group is able to displace both the TIP3P and the TIP4P water within 100 ns. In the MD simulation with TIP3P water, the water displacement occurs at 53 ns, while in the simulation with TIP4P it occurs at 50 ns. The OH group of the hydroxyethyl forms stable hydrogen bonds to the three binding site key residues

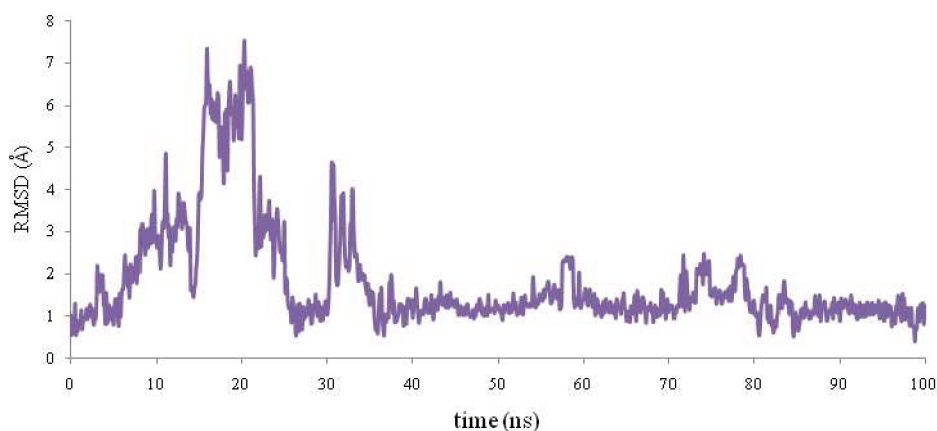


Figure 8. RMSD values for the water oxygen calculated over 100 ns of unrestrained MD simulations of the Con A–3HET complex in TIP4P water.

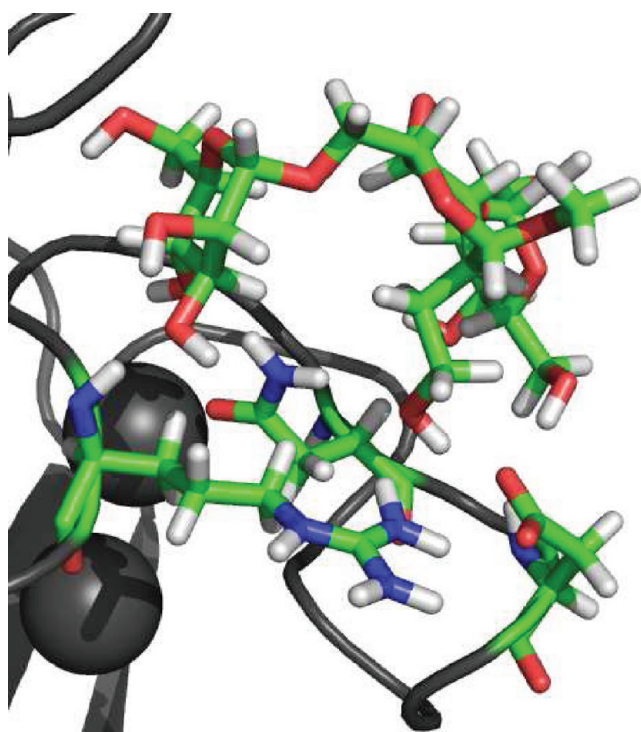


Figure 9. Hydroxyethyl of 3HET occupying Con A's binding site after the displacement of the conserved water. The conformation was obtained from the 100 ns of unrestrained MD simulation of the Con A–3HET complex in TIP3P water.

(see Figure 9), and it occupies the location of the conserved water for the remainder of the simulation. The average hydrogen bond distances are shown in Table 8. A comparison between the data in Tables 7 and 8 shows that the hydroxyethyl OH is coordinated exactly like the conserved water in both the simulation with TIP3P and that with TIP4P.

As seen in the simulation of the Con A–3MAN complex with TIP5P, in the simulation of the Con A–3HET complex with TIP5P, no exchanges between conserved water and the bulk were observed. Additionally, within the 100 ns of unrestrained simulation, the hydroxyethyl group was unable to displace the conserved water (see Table 7 for hydrogen bond distances).

Table 7. Hydrogen Bond Distances^a between the Conserved Water and the Key Residues in the Binding Site of the Con A–3HET Complex

H-bond distances	3D4K	TIP3P	TIP4P	TIP5P
O _{WAT} –O _{ASP16}	2.5	2.7 (0.3) ^b	3.0 (0.6)	2.7 (0.1)
O _{WAT} –N _{ASN14}	2.7	3.0 (0.2)	3.0 (0.2)	2.8 (0.1)
O _{WAT} –N _{ARG228}	3.0	3.0 (0.3)	3.0 (0.4)	2.9 (0.2)
O _{WAT} –O _{2HET}	3.0	3.1 (0.5)	3.3 (0.9)	4.3 (0.9)

^a In Ångstroms. ^b Standard deviations are indicated in parentheses.

Table 8. Hydrogen Bond Distances between^a the Terminal OH of the Hydroxyethyl Group of 3HET and the Key Residues in the Binding Site of the Con A–3HET Complex^c

H-bond distances	TIP3P	TIP4P
OH _{3HET} –O _{ASP16}	2.7 (0.2) ^b	2.8 (0.3)
OH _{3HET} –N _{ASN14}	3.0 (0.2)	3.0 (0.2)
OH _{3HET} –N _{ARG228}	3.0 (0.2)	3.0 (0.2)

^a In Ångstroms. ^b Standard deviations are indicated in parentheses.

^c Average distances are calculated during the unrestrained MD simulations in TIP3P and TIP4P water only after the conserved water is displaced. TIP5P water was not displaced within the 100 ns trajectory.

CONCLUSIONS

In this work, we analyzed the function and stability of a structurally conserved water molecule in the binding site of the carbohydrate binding protein Con A by means of binding free energy calculations and MD simulations. The calculations were performed for three systems: the free protein, the protein bound to its natural ligand (3MAN), and the protein bound to a synthetic analog (3HET) that was designed and synthesized in previous work to displace the conserved water.^{24,25} Subsequent structural data²⁵ showed that 3HET was not in fact successful in displacing the key water. To provide insight into the unexpected resistance of this water to displacement, its binding affinity was estimated employing three rigid nonpolarizable water models, namely, TIP3P, TIP4P, and TIP5P. The water binding free energy values, the dynamics, and the stability of the water in the binding site were all found to depend significantly on the choice of water model. On the basis of the free energy values and the dynamics, in the free protein, TIP3P clearly underestimates the

level of order and occupancy of the binding site by the conserved water, while both TIP3P and TIP4P correctly predict that the binding site is specifically hydrated.

In the complex with 3MAN, the TIP3P water is tightly bound and optimally coordinated with the three key amino acid residues in the Con A binding site and the hydroxyl group of the central mannose. In contrast, both TIP3P and TIP4P are less strongly bound and more labile during the simulations.

In the Con A–3HET complex, both TIP3P and TIP4P escape from the binding site to be replaced by the hydroxyl group of the hydroxyethyl moiety. This direct interaction between the synthetic ligand and the Con A binding site was predicted from earlier studies with TIP3P²³ but was refuted by subsequent crystallographic data.²⁵ Only the TIP3P model maintains its position in the 3HET complex throughout the 100 ns of simulation, although it is less strongly bound than in the complex with the natural ligand.

Although the MD simulations with TIP3P and TIP4P suggest that the displacement of the conserved water by a synthetic ligand is an easier task than it apparently is, the length of the hydroxyethyl group does indeed fit this requirement. However, the flexibility of the hydroxyethyl moiety in 3HET permits it to adopt a conformation that enables 3HET to bind without displacing the water. A modification of the hydroxyethyl group that enhanced the stability of the C2–C7 torsion in a g^+ conformation could conceivably displace this water, provided that the net ligand binding free energy exceeded -2.4 kcal mol⁻¹, the binding free energy computed with TIP3P for the conserved water in free Con A.

■ ASSOCIATED CONTENT

S Supporting Information. The results obtained with the TIP3P-MOD water model together with additional figures and tables have been included. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +1 (706) 542 4454. Fax: +1 (706) 542 4412. E-mail: rwoods@ccrc.uga.edu

■ ACKNOWLEDGMENT

E.F. thanks Chris Neale and Régis Pomès for insightful discussions. The Science Foundation of Ireland (SFI operating grant 08/IN.1/B2070) is gratefully acknowledged for funding.

■ REFERENCES

- (1) Dwek, R. A. *Chem. Rev.* **1996**, *96*, 683–720.
- (2) Rini, J. M. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 551–577.
- (3) Lis, H.; Sharon, N. *Chem. Rev.* **1998**, *98*, 637–674.
- (4) de Beer, S. B. A.; Vermeulen, N. P. E.; Oostenbrink, C. *Curr. Top. Med. Chem.* **2009**, *10*, 55–66.
- (5) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2009**, *131*, 15403–15411.
- (6) Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 500–512.
- (7) Homans, S. W. *Drug Discovery Today* **2007**, *12*, 534–539.
- (8) Baron, R.; Setny, P.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 12091–12097.

- (9) Setny, P.; Baron, R.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 2866–2871.
- (10) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
- (11) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587.
- (12) Gupta, D.; Dam, T. K.; Oscarson, S.; Brewer, C. F. *J. Biol. Chem.* **1997**, *272*, 6388–6392.
- (13) Brewer, C. F.; Bhattacharyya, L. *J. Biol. Chem.* **1986**, *261*, 7306–7310.
- (14) Naismith, J. H.; Field, R. A. *J. Biol. Chem.* **1996**, *271*, 972–976.
- (15) Bhattacharyya, L.; Brewer, C. F. *Eur. J. Biochem.* **1989**, *178*, 721–726.
- (16) Chervenak, M.; Toone, E. J. *Biochemistry* **1995**, *34*, 5685–5695.
- (17) Bryce, R. A.; Hillier, I. H.; Naismith, J. H. *Biophys. J.* **2001**, *81*, 1373–1388.
- (18) Bouckaert, J.; Hamelryck, T. W.; Wyns, L.; Loris, R. *J. Biol. Chem.* **1999**, *274*, 29188–29195.
- (19) Dam, T. K.; Oscarson, S.; Sacchetti, J. C.; Brewer, C. F. *J. Biol. Chem.* **1998**, *273*, 32826–32.
- (20) Mandal, D. K.; Kishore, N.; Brewer, C. F. *Biochemistry* **1994**, *33*, 1149–1156.
- (21) Katardjieff, K.; Höchtl, P.; Segelke, B. W.; Tao, F.-M.; Rupp, B. *Acta Crystallogr.* **2002**, *58*, 735–743.
- (22) Deacon, A.; Gleichmann, T.; Kalb (Gilboa), A. J.; Price, H.; Raftery, J.; Bradbrook, G.; Yariv, J.; Helliwell, J. R. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 4305–4312.
- (23) Parkin, S.; Rupp, B.; Hope, H. *Acta Crystallogr.* **1996**, *DS2*, 1161–1168.
- (24) Clarke, C.; Woods, R. J.; Gluska, J.; Cooper, A.; Nutley, M. A.; Boons, G. J. *J. Am. Chem. Soc.* **2001**, *123*, 12238–12247.
- (25) Kadirvelraj, R.; Foley, B. L.; Dyekjaer, J. D.; Woods, R. J. *J. Am. Chem. Soc.* **2008**, *130*, 16933–16942.
- (26) Li, Z.; Lazaridis, T. *J. Phys. Chem. B* **2005**, *109*, 662–670.
- (27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (28) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.
- (29) Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3569–3579.
- (30) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
- (31) McKenzie, G. H.; Sawyer, W. H. *J. Biol. Chem.* **1973**, *248*, 549–556.
- (32) Sun, Y. X.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1164–1169.
- (33) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- (34) York, D. M.; Darden, T. A.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *99*, 8345–8348.
- (35) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (36) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.
- (37) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622–655.
- (38) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 13.
- (39) Olano, L. R.; Rick, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 7991–8000.
- (40) Kantardjieff, K. A.; Hochtl, P.; Segelke, B. W.; Tao, F. M.; Rupp, B. *Acta Crystallogr.* **2002**, *58*, 735–743.

Do Two Different Reaction Mechanisms Contribute to the Hydroxylation of Primary Amines by Cytochrome P450?

Patrik Rydberg* and Lars Olsen

Department of Medicinal Chemistry, Copenhagen University, Universitetsparken 2, DK-2100 Copenhagen, Denmark

Supporting Information

ABSTRACT: Three possible mechanisms have been suggested for the hydroxylation of primary and secondary amines by the cytochrome P450 enzyme family. We show that for the hydroxylation of primary alkyl amines, both the hydrogen abstraction and rebound mechanism and the direct oxygen transfer mechanism can contribute to the formation of the hydroxylated product. We also show that in the hydrogen abstraction and rebound mechanism the rebound step has higher activation energy than the hydrogen abstraction step, which is the opposite of the hydroxylation of aliphatic carbon atoms.

INTRODUCTION

Many primary amines have been shown to undergo amine hydroxylation in cytochromes P450.¹ Hence, knowledge of the hydroxylation mechanism and the rate limiting step is of vital importance for reactivity-based models that predict CYP mediated drug metabolism. Our model substrate, propan-2-amine, is a fragment of amphetamine and mexiletine, which both undergo amine hydroxylation.^{2,3} The hydroxylation of primary and secondary amines is also the first step in the formation of nitrones,⁴ which can cause the inhibition of cytochromes P450s by the formation of a bond between the nitron nitrogen atom and the heme iron atom.⁵

During the past decade, the reaction mechanisms of reactions performed by the cytochrome P450 (CYP) enzyme family have been clarified by experimental and quantum chemical studies one by one,^{6,7} but there are still many mechanisms that have not been studied in great depth. In this study, we investigate the possible mechanisms for hydroxylation of primary amines using density functional theory (DFT) for the first time.

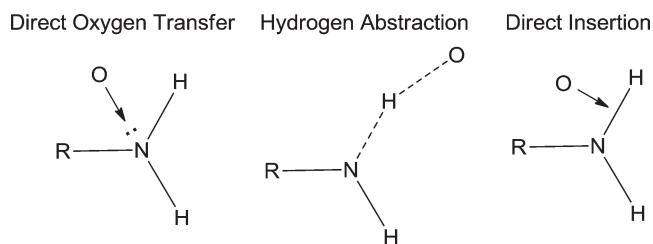
Three possible reaction mechanisms have been suggested for the hydroxylation of primary and secondary amines,⁸ and these are shown in Scheme 1: direct oxygen transfer (addition of the oxygen to the nitrogen lone pair followed by a rearrangement of the formed N-oxide into the hydroxylamine), hydrogen abstraction from the nitrogen followed by a rebound step, and direct insertion of the oxygen into the N–H bond. We have investigated the three mechanisms by performing DFT calculations on a porphine model system with propan-2-amine as a substrate.

COMPUTATIONAL METHODOLOGY

All calculations were performed with the Turbomole software package,⁹ version 6.1. In the calculations, compound I in the CYPs is modeled by a reduced heme model without side chains, iron porphine with SCH_3^- , and O^{2-} as axial ligands.

All calculations were performed using the B3LYP functional^{10–12} within the VWN(V) correlation functional¹³ (unrestricted formalism for open shell systems). The geometry optimizations, frequency calculations, and solvent calculations were performed

Scheme 1. The Three Mechanisms Suggested for the Hydroxylation of Primary and Secondary Amines



with the double- ζ basis set of Schäfer et al.,¹⁴ enhanced with a p function with the exponent 0.134915, on the iron atom, and the 6-31G(d) basis set^{15–17} for the other atoms. The final energies were determined by single point calculations using the 6-311++G(2d,2p) basis set^{18,19} for all atoms, except iron, for which we used the double- ζ basis set of Schäfer et al.,¹⁴ enhanced with s , p , d , and f functions (exponents of 0.01377232, 0.041843, 0.1244, 2.5, and 0.8; two f functions).²⁰

Solvent calculations were carried out with the continuum conductor-like screening model (COSMO),²¹ using an effective dielectric constant (ϵ) of 4 (except where otherwise mentioned). For the atomic radii, we used the optimized COSMO radii in Turbomole⁹ (and 2.0 Å for Fe).

All energies presented have been computed using the large basis set and include zero point vibrational corrections and solvation effects except where otherwise mentioned.

The barrier for the transformation of the direct oxygen transfer product (P^{Ox}) into the hydroxylated product (P^{Ha}) was computed using a propan-2-amine molecule and three water molecules (two of the water molecules created a proton transfer chain moving a proton from the nitrogen atom to the oxygen atom).

The hydrogen bond strength of nitrogen and carbon radicals to the hydroxyl group in the intermediate in the hydrogen

Received: June 20, 2011

Published: September 13, 2011

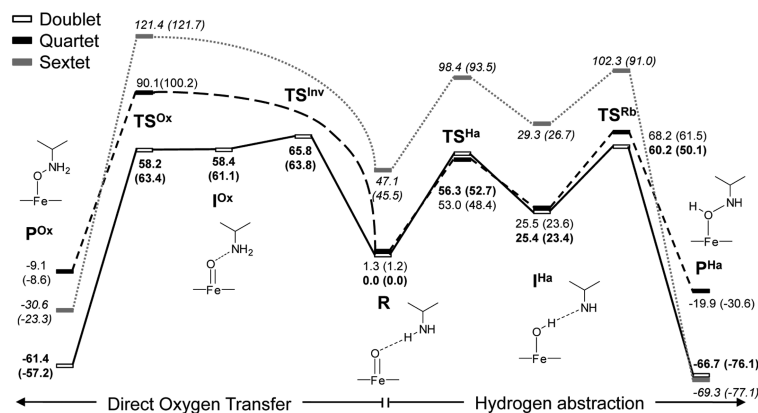


Figure 1. Energies for the direct oxygen transfer and hydrogen abstraction mechanisms for the doublet (bold text), quartet, and sextet (italic text) spin states in kilojoules per mole. Energies are from single-point calculations with the large basis set and include zero-point vibrational energies and solvent corrections computed with a dielectric constant of 4. The energies in parentheses are vacuum energies without solvent corrections.

abstraction mechanism were computed using model systems consisting of a water molecule and either an NH_2 radical or a CH_3 radical, with the 6-31G(d) basis set.^{15–17}

Spin densities were computed by Mulliken population analysis of the single-point calculations of the large basis set combination.

RESULTS AND DISCUSSION

The direct insertion mechanism has been shown to be unlikely for primary alkylamines,²² and later work has suggested that this mechanism should be more probable for nitrogen atoms with a highly delocalized lone pair.⁸ Our results support that the insertion mechanism is very unlikely for primary alkyl amines. Extensive work was done to find the transition state of this mechanism, starting from several different structures and using different reaction coordinates. However, each time we tried to locate the transition state for the insertion mechanism, the optimization ended up in structures belonging to either the direct oxygen transfer mechanism or the hydrogen abstraction mechanism. Hence, this mechanism will not be discussed any further.

The direct oxygen transfer mechanism has previously been studied for tertiary amines in the doublet and quartet spin states.^{23–26} In this work, we also study the mechanism in the sextet spin state. The transition state of the oxidation (TS^{Ox}) is quite similar to the results for the tertiary amine trimethylamine,²³ with a high barrier of 90 kJ/mol for the quartet spin state and a lower barrier of 58 kJ/mol for the doublet spin state (see Figure 1). In the sextet spin state, the barrier is even larger than in the quartet spin state (121 kJ/mol). This large energetic difference occurs because in the doublet spin state the two accepting orbitals to which the electrons from the nitrogen lone pair are transferred (one on the oxy group and one shared by the porphyrin ring and the cysteine sulfur atom) are singly occupied by electrons with opposite spin, and hence they can accept both the electrons in a straightforward manner. However, in the quartet (and sextet) spin states these orbitals are occupied by electrons with the same spin, and one of the electrons must either flip spin or end up in another orbital (an unoccupied iron 3d orbital), as has been shown previously for both nitrogen oxidation²³ and sulfur oxidations.^{7,23} While the spin density distributions basically are the same in propan-2-amin and trimethylamine, the structures have one difference:

the iron–oxygen–amine nitrogen angle is smaller for propan-2-amine due to weak interactions between an amine hydrogen atom and two of the nitrogen atoms in the porphine ring (structures and spin densities of the direct oxidation mechanism are shown in Figure 2). The major difference is that for propan-2-amine there is a requirement for an additional inversion step to get from the reactant state (R) to the direct oxidation transition state (TS^{Ox}). This step is required to move the nitrogen lone pair toward the oxygen, since it points away from the iron-bound oxygen atom in the reactant state (R) where there is a hydrogen bond between the iron-bound oxygen atom and one of the amine hydrogen atoms. The transition state of this nitrogen inversion (TS^{Inv}) is quite similar to the following intermediate and oxidation transition state with regard to spin densities (see Figure 2). However, the normal mode of the imaginary vibration is almost a pure nitrogen inversion in which the nitrogen atom is moving toward the oxygen atom (the structure distorted along the normal mode is shown in Figure S1, Supporting Information). While the intermediate (I^{Ox}) in the doublet state has a slightly higher energy than the transition state (0.2 kJ/mol), this is only due to solvation effects, because without these the intermediate (I^{Ox}) is lower in energy by 2.3 kJ/mol. Stable structures of this intermediate and the inversion transition state do not exist in the quartet or sextet spin states. This seems to be due to the high barrier of the direct oxidation in these spin states; what should have been the inversion transition state (TS^{Inv}) only shows up as a shoulder on scans of the oxygen nitrogen distance, because at this iron–oxygen distance the energy of the direct oxidation process is already quite high.

The rearrangement of the product of the direct oxygen transfer into the hydroxylated nitrogen (the product of the hydrogen abstraction and rebound reaction) in water has a lower barrier than either of the two reaction paths (18.3 kJ/mol). Hence, the direct oxidation product will be rearranged into the hydroxylated product more rapidly than it is formed, making the product of the direct oxidation unlikely to be observed in experiments.

The hydrogen abstraction and rebound mechanism has previously been studied extensively for the hydroxylation of aliphatic carbon atoms,⁶ and the hydrogen abstraction from the hydroxyl oxygen in ethanol has also been studied.²⁷ Since the hydrogen abstraction from nitrogen should be similar to the hydrogen abstraction from oxygen, we compare our results primarily to the

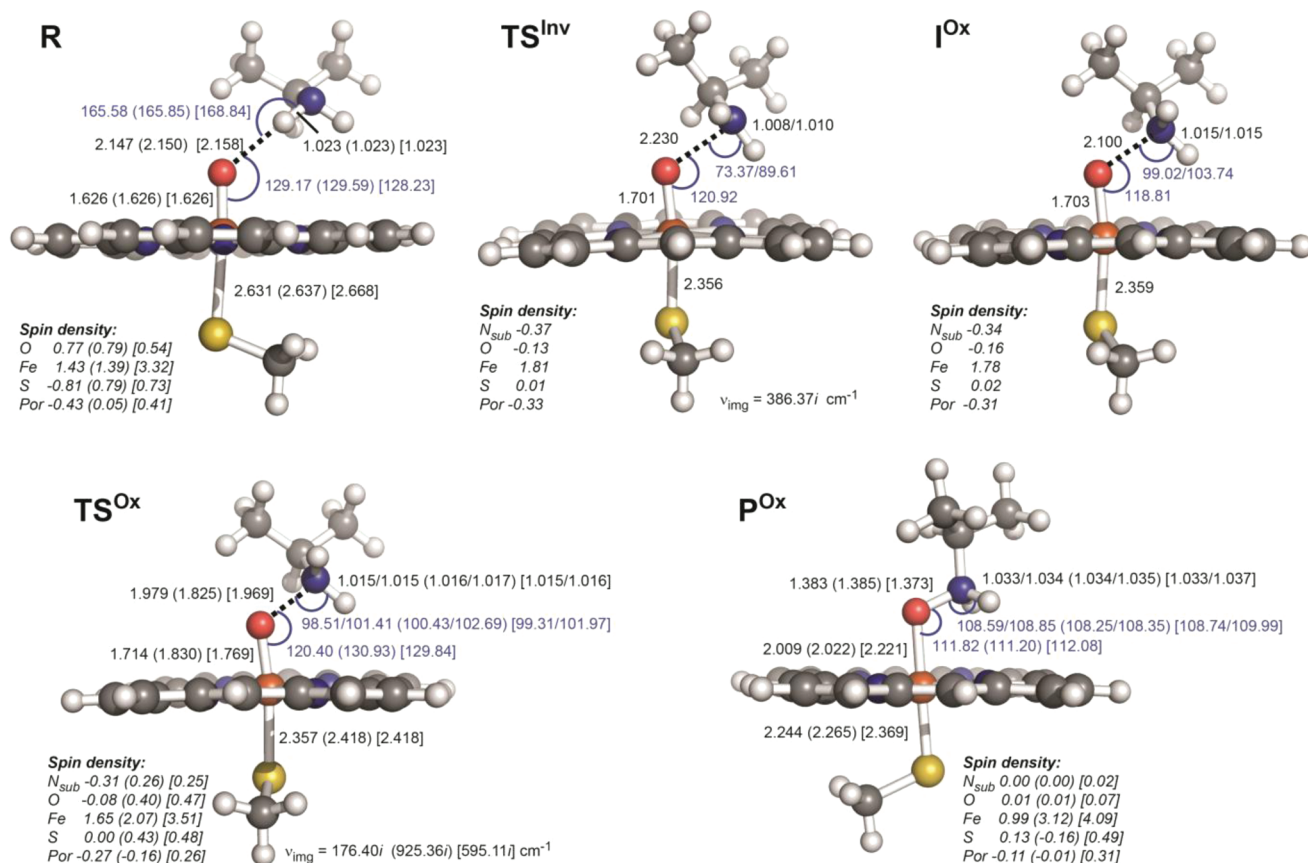


Figure 2. Structures, imaginary frequencies, geometrical features, and spin densities for the direct oxidation pathway. All data are shown as doublet [quartet] [sextet]. Distances in Ångströms are shown in black, and angles in degrees are shown in blue.

ones of Wang et al.,²⁷ who studied hydrogen abstraction from both oxygen and carbon atoms in ethanol.

The different states in the hydrogen abstraction and rebound mechanism are in general structurally similar to the results on ethanol,²⁷ but due to the formation of hydrogen bonds of different strengths by oxygen and nitrogen atoms compared to a carbon atom there are some significant differences. In the doublet spin state, the hydrogen abstraction transition state (TS^{Ha}) in propan-2-amine has an O–H distance which is similar to that for the hydrogen abstraction from carbon but longer than the one for hydrogen abstraction from oxygen, while the H–N distance is ~ 0.1 Å shorter than the same H–C/O distances. In the quartet spin state, the O–H distance is intermediate between the corresponding distances in C/O hydrogen abstraction. Energetically, the hydrogen abstraction barrier is similar for all three reactions in both doublet and quartet spin states (49–56 kJ/mol). Since the study by Wang et al.²⁷ did not use the same model system and basis set as we do in this study, the differences in energies are most likely not significant. Earlier work has shown that changing the model system can change the activation energies of P450 mediated reactions by roughly 5 kJ/mol (our SCH_3^- model giving higher energies), but geometrically, the only significant change is in the iron–sulfur bond distance.²³ The sextet spin state is geometrically quite similar to the quartet spin state, but energetically it is much higher throughout the abstraction and rebound mechanism (with the exception of the product state). For the optimization of the intermediate structure (I^{Ha}) in the sextet spin state, the iron–sulfur bond had to be constrained to

generate a stable structure. The intermediates (I^{Ha}) are also structurally very similar with the exception of the distance between the oxygen bound hydrogen atom and the amine nitrogen radical (1.9 Å, see Figure 3) which is intermediate between the corresponding distances in the hydrogen abstraction from oxygen (1.8 Å) and carbon (2.0 Å) atoms.²⁷ This strong nitrogen–hydrogen interaction causes the following rebound step to be quite different in the amine hydroxylation compared to the same step in the hydroxylation of aliphatic carbon atoms. Whereas the largest activation energy for the rebound step in the hydroxylation of the aliphatic carbons is only 13 kJ/mol higher than the intermediate (quartet spin state in the hydroxylation of camphor),²⁸ the lowest corresponding energy for the amine hydroxylation is 35 kJ/mol (TS^{Rb} , doublet spin state). This is also evident from the structure of the rebound transition state which shows a much later transition state with a shorter O–N distance compared to the corresponding transition state in the aliphatic carbon hydroxylation in camphor.²⁸ The reason for this high barrier in the rebound step is that the mechanism is actually slightly different compared to the rebound step in the hydroxylation of aliphatic carbon atoms. In the hydroxylation of aliphatic carbon atoms, the intermediate step (I^{Ha}) with a carbon radical has the radical oriented toward the hydroxyl group, and there is only one possible orientation. However, in the hydroxylation of nitrogen atoms, the radical on the nitrogen atom is actually almost perpendicular to the direction of the hydroxyl group, and the lone pair is interacting with the hydroxyl hydrogen atom, as shown in Scheme 2. This different orientation forces the

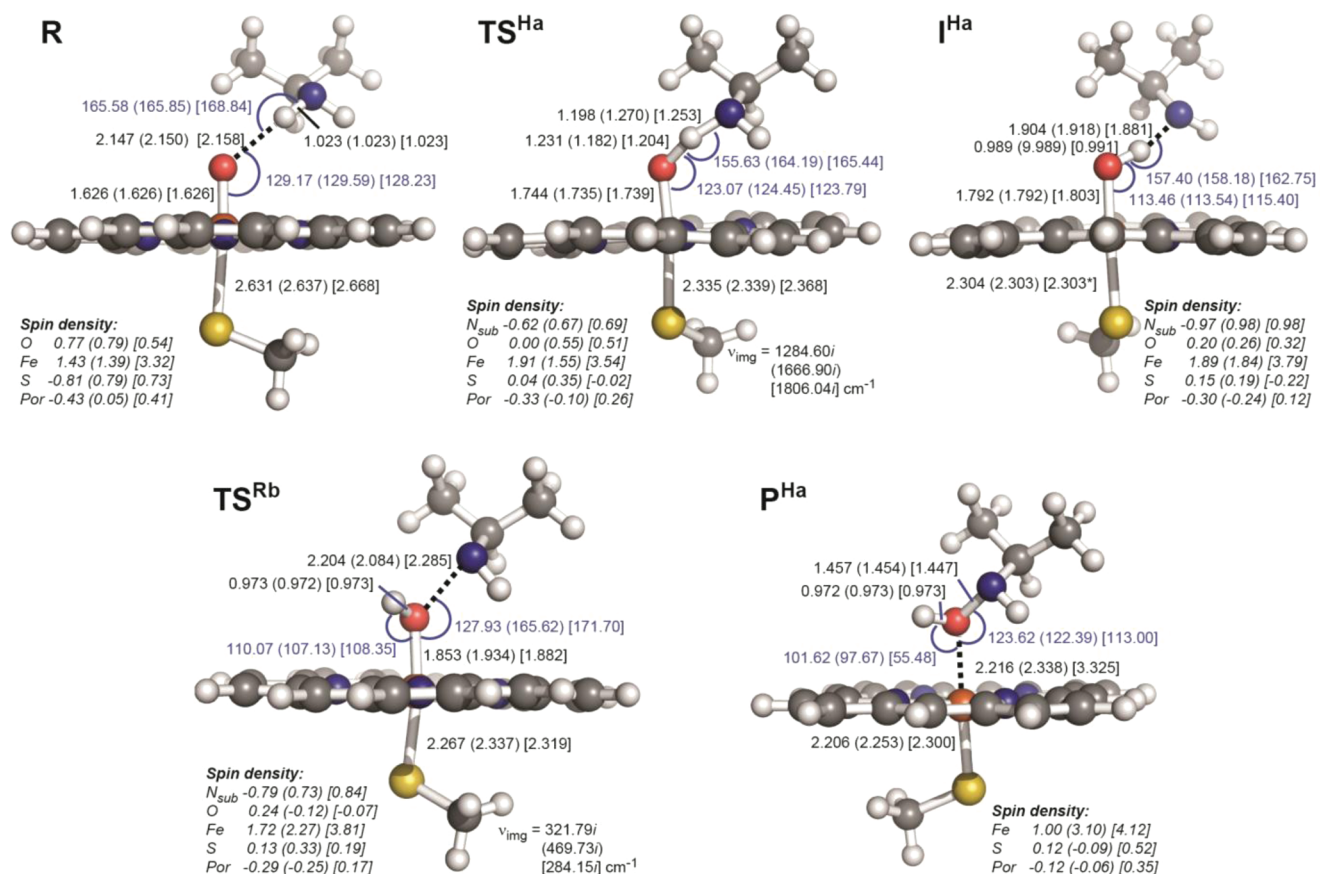
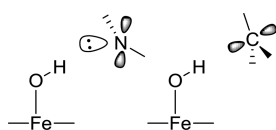


Figure 3. Structures, imaginary frequencies, geometrical features, and spin densities for the hydrogen abstraction and rebound pathway. All data are shown as doublet (quartet) [sextet]. Distances in Ångstroms shown in black and angles in degrees shown in blue. The distance that was constrained during geometry optimization of the intermediate (I^{Ha}) in the sextet spin state is labeled with *. The reactant state (R) is identical to the one in Figure 2.

Scheme 2. The Different Radical Distributions in the Hydrogen Abstraction Intermediate for Amines and Aliphatic Carbons



system into an electronic reorganization during the rebound step, since the radical and the lone pair of the nitrogen have to change places. There is also a significant difference in hydrogen bond strength of the carbon and nitrogen radicals, which is another reason for the higher rebound barrier. The nitrogen radical has a 16 kJ/mol stronger hydrogen bond ($N\cdots H-O$) than the corresponding bond for a carbon radical. This hydrogen bond has to be broken during the rebound since the hydroxyl group has to be rotated to allow the oxygen lone pair to bind to the amine nitrogen. The barrier of the rebound transition state (TS^{Rb}) in the amine hydroxylation has an energy that actually is higher than the one for the hydrogen abstraction transition state (TS^{Ha}) for all three spin states. This is quite different from the same mechanism in the hydroxylation of aliphatic carbon atoms, where the doublet spin state rarely has a rebound barrier, while it is more common in the quartet spin state.²⁹ Still, even when there is a barrier for the rebound of aliphatic carbons, its

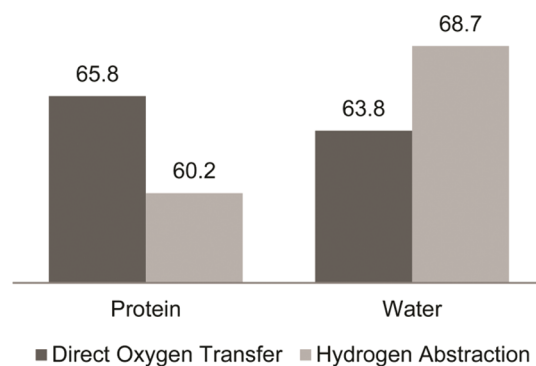


Figure 4. Comparison of the highest energies of the direct oxygen transfer and hydrogen abstraction mechanisms for solvation effects computed with water ($\epsilon = 80$) or in a protein ($\epsilon = 4$; energies in kJ/mol). The additional solvation effects of water make the direct oxygen transfer mechanism more favorable than the hydrogen abstraction mechanism.

energy is always much lower than the one for the hydrogen abstraction.

While the reaction energies show that the hydrogen abstraction mechanism is slightly more favorable than the direct oxygen transfer mechanism, this preference changes if we mimic water solvation by increasing the dielectric constant in our implicit

solvent model from 4 to 80 (see Figure 4). This indicates that explicit hydrogen bonds to the amine could change the preference to a direct oxygen transfer mechanism.

To investigate whether interactions with amino acids or water molecules are likely to contribute to the *N*-hydroxylation of primary amines, we investigated how the amine group in mexiletine interacts with amino acids in the active site of CYP1A2 when it is positioned in a way favorable for *N*-hydroxylation. To do this, we analyzed our docking results from a previous study,³⁰ and it is clear that there are no direct interactions between the amine group and any amino acids. Hence, interactions with water molecules which could exist in the active site would be required to shift the balance between the two mechanisms for the amine hydroxylation of mexiletine by CYP1A2.

CONCLUSIONS

The results of this study show that hydroxylation of primary alkyl amines (and probably also secondary alkyl amines) could undergo hydroxylation by cytochromes P450 either through a hydrogen abstraction and rebound mechanism or through the direct oxygen transfer mechanism, depending on the interactions between the substrate and water molecules or amino acids in the active site in each case.

There is one major difference in the hydrogen abstraction and rebound mechanism for hydroxylation of primary amines compared to the previously extensively studied hydroxylation of aliphatic carbon atoms, and that is the rebound step. Our results show that the radical intermediate has a different radical distribution in primary amines compared to aliphatic carbon atoms, resulting in a much stronger hydrogen bond in the intermediate and a much higher barrier for the rebound step. The barrier for the rebound step is even higher than the one for the hydrogen abstraction step in the hydroxylation of propan-2-amine.

ASSOCIATED CONTENT

S Supporting Information. Coordinates for all structures mentioned in the text. Full data for the rearrangement of the product of direct oxygen transfer into the *N*-hydroxylated product. Tables with energies for all structures. Figure of the distortion along the normal mode for the inversion transition state. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: (+45) 35 33 61 62. Fax: (+45) 35 30 60 40. E-mail: pry@farma.ku.dk

ACKNOWLEDGMENT

This work was supported by grants from the Danish medical research council and Lhasa Limited.

REFERENCES

- (1) Hlavica, P. *N*-oxidative transformation of free and *N*-substituted amine functions by cytochrome P450 as means of bioactivation and detoxication. *Drug Metab. Rev.* **2002**, *34*, 451–477.
- (2) Florence, V. M.; Distefano, E. W.; Sum, C. Y.; Cho, A. K. The Metabolism of (R)-(-)-Amphetamine by Rabbit Liver-Microsomes - Initial Products. *Drug Metab. Dispos.* **1982**, *10*, 312–315.

- (3) Labbe, L.; Abolfathi, Z.; Lessard, E.; Pakdel, H.; Beaune, P.; Turgeon, J. Role of specific cytochrome P450 enzymes in the *N*-oxidation of the antiarrhythmic agent mexiletine. *Xenobiotica* **2003**, *33*, 13–25.
- (4) Bondon, A.; Macdonald, T. L.; Harris, T. M.; Guengerich, F. P. Oxidation of Cycloalkylamines by Cytochrome-P450 - Mechanism-Based Inactivation, Adduct Formation, Ring Expansion, and Nitrene Formation. *J. Biol. Chem.* **1989**, *264*, 1988–1997.
- (5) Murray, M. Drug-mediated inactivation of cytochrome P450. *Clin. Exp. Pharmacol. Physiol.* **1997**, *24*, 465–470.
- (6) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. *Chem. Rev.* **2005**, *105*, 2279–2328.
- (7) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. P450 Enzymes: Their Structure, Reactivity, and Selectivity-Modeled by QM/MM Calculations. *Chem. Rev.* **2010**, *110*, 949–1017.
- (8) Nishida, C. R.; Knudsen, G.; Straub, W.; de Montellano, P. R. O. Electron supply and catalytic oxidation of nitrogen by cytochrome P450 and nitric oxide synthase. *Drug Metab. Rev.* **2002**, *34*, 479–501.
- (9) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J. C. W. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.
- (10) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (11) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula Into A Functional of the Electron-Density. *Phys. Rev. B* **1988**, *37*, 785–789.
- (12) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (13) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin-Density Calculations - A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (14) Schafer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian-Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (15) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular-Orbital Studies of Organic-Molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
- (16) Hariharan, P. C.; Pople, J. A. Influence of Polarization Functions on Molecular-Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (17) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; Defrees, D. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77*, 3654–3665.
- (18) Mclean, A. D.; Chandler, G. S. Contracted Gaussian-Basis Sets for Molecular Calculations. I. Second Row Atoms, Z=11–18. *J. Chem. Phys.* **1980**, *72*, 5639–5648.
- (19) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. XX. Basis Set for Correlated Wave-Functions. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (20) Rulisek, L.; Jensen, K. P.; Lundgren, K.; Ryde, U. The reaction mechanism of iron and manganese superoxide dismutases studied by theoretical calculations. *J. Comput. Chem.* **2006**, *27*, 1398–1414.
- (21) Klamt, A.; Schuurmann, G. Cosmo - A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.
- (22) Pack, G. R.; Loew, G. H. Semiempirical studies of the mechanism of models for the *N*-hydroxylation of amines by cytochrome P450. *Int. J. Quantum Chem.* **1979**, *16*, 381–390.
- (23) Rydberg, P.; Ryde, U.; Olsen, L. Sulfoxide, sulfur, and nitrogen oxidation and dealkylation by cytochrome P450. *J. Chem. Theory Comput.* **2008**, *4*, 1369–1377.
- (24) Li, C. S.; Wu, W.; Cho, K. B.; Shaik, S. Oxidation of Tertiary Amines by Cytochrome P450-Kinetic Isotope Effect as a Spin-State Reactivity Probe. *Chem.—Eur. J.* **2009**, *15*, 8492–8503.
- (25) Cho, K. B.; Moreau, Y.; Kumar, D.; Rock, D. A.; Jones, J. P.; Shaik, S. Formation of the active species of cytochrome P450 by using

iodosylbenzene: A case for spin-selective reactivity. *Chem.—Eur. J.* **2007**, *13*, 4103–4115.

(26) Roberts, K. M.; Jones, J. P. Anilinic N-Oxides Support Cytochrome P450-Mediated N-Dealkylation through Hydrogen-Atom Transfer. *Chem.—Eur. J.* **2010**, *16*, 8096–8107.

(27) Wang, Y.; Yang, C. L.; Wang, H. M.; Han, K. L.; Shaik, S. A new mechanism for ethanol oxidation mediated by cytochrome P450 2E1: Bulk polarity of the active site makes a difference. *ChemBiochem* **2007**, *8*, 277–281.

(28) Kamachi, T.; Yoshizawa, K. A theoretical study on the mechanism of camphor hydroxylation by compound I of cytochrome P450. *J. Am. Chem. Soc.* **2003**, *125*, 4652–4661.

(29) Shaik, S.; Hirao, H.; Kumar, D. Reactivity patterns of cytochrome P450 enzymes: multifunctionality of the active species, and the two states-two oxidants conundrum. *Nat. Prod. Rep.* **2007**, *24*, 533–552.

(30) Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L. Fast Prediction of Cytochrome P450 Mediated Drug Metabolism. *ChemMedChem* **2009**, *4*, 2070–2079.

Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies

Jeffrey K. Weber and Vijay S. Pande*

Department of Chemistry, Stanford University, Stanford, California 94305, United States

ABSTRACT: Markov state models (MSMs) have proven themselves to be effective statistical and quantitative models for understanding protein folding dynamics. As stochastic networks, MSMs allow for descriptions of parallel folding pathways and facilitate quantitative comparison to experiments conducted at the ensemble level. While this complex network structure is advantageous in many respects, a simple topological description of these graphs is elusive. In this Article, we compare a series of protein folding MSMs to the topology of the Cayley tree, a graph structure on which dynamics are intuitive. We go on to introduce and test new sampling schemes that have potential to improve automated model construction, a critical step toward making Markov state modeling more accessible to general users.

INTRODUCTION

Simulations of biological polymers have advanced from being simple depictions of dynamics to providing statistical and quantitative descriptions of the self-assembly process.^{1–4} In protein folding in particular, efforts to create statistically grounded models have been focused on a discrete master equation approach called the Markov state model (MSM).^{5,6} MSMs take advantage of parallel sampling techniques by partitioning a protein's configuration space into a set of kinetically distinct states. Upon determining the time scale on which transitions between these states are memoryless, an MSM transition matrix can advance dynamics to the long time scales necessary to describe folding processes. Recent millisecond time scale simulations of the protein NTL9 and the five-helix bundle of λ -repressor show the promise of MSMs in simulating slowly folding systems.^{7,8}

As quantitative comparison of simulation with experiment becomes not only desirable but imperative, MSMs offer a convenient avenue for modeling protein folding on an ensemble level. The extensive theory of Markov chains allows kinetic and equilibrium properties for the ensemble to be easily extracted from the eigenspectrum of a transition matrix. The stationary distribution vector (the eigenvector with unit eigenvalue) describes state population probabilities at equilibrium. A protein's native state can be identified from this equilibrium distribution without a priori knowledge of structure, simply by noting the state with the highest stationary population. The other eigenvectors of the transition matrix describe dynamical processes at time scales determined by their eigenvalues, allowing one to deduce which states are kinetically relevant over short and long time periods. Other ensemble properties like the mean first-passage time to the native state can also be quickly calculated from well-known statistical theory.⁶

A description of protein folding under a conventional two-state folding model is intuitive: molecules proceed from “unfolded” to “folded” in a concerted matter, and the “rate of folding” is well-defined by the transition between these two states. MSMs, however, describe dynamics on a network of many hundreds or thousands of states that are connected by probability-weighted edges. It is not immediately clear which states should be called

“unfolded” or “intermediate” states, or which correspond to the most biologically relevant structures. Folding rates to the native state are well-defined from all of these states and can be highly disparate. Analysis of network connectivity (involving degree and distance from the native state) is necessary to both classify states and to make quantitative kinetic predictions. As previous connectivity-based analysis has been performed on an ad hoc basis, a general description of protein folding network topology would be of interest.

With such a general description of connectivity, one could also tailor MD sampling strategies for MSM topologies. Given the increasing popularity of Markov models in biomolecular simulation, it is of general interest to make MSMs more accessible to nonexpert users. Recent projects like MSMBuild2 and Copernicus have made strides in automating the construction of MSMs from raw molecular dynamics data and (in the case of Copernicus) even more general user-defined protocol.^{10,11} Instrumental to this automation has been the development of “adaptive sampling”, which actively pushes simulations toward under-sampled regions of configuration space.^{12–14} Specifically, adaptive sampling starts trajectories from states that contribute the maximum uncertainty to the model's largest nonunit eigenvalue. This adaptation prevents the simulation from being stuck in metastable free energy wells for untenably long periods of wall-clock time, a critical procedure for ensuring sampling efficiency and model refinement.

As the fine details of automatic model construction become better understood, however, the utility of using eigenvalue-based sampling early on in the process has come into question. Particularly, the model's state decomposition, which current adaptive sampling schemes presume to be finalized, is itself subject to a high degree of uncertainty in early stages of sampling.¹¹ Refining a model based on a poor partitioning will naturally reduce the effectiveness of eigenvalue-based sampling. At present, intermittent rounds of randomly distributed trajectories are prescribed to address this problem.

Received: June 29, 2011

Published: September 07, 2011

Can a more systematic strategy be devised for early model refinement? The strategy extended in this study is founded on an “adjacency-based” sampling scheme, focusing on determining the connectivity of the transition matrix. A model’s adjacency matrix defines many of its fundamental characteristics. An observed transition between two states indicates that the barrier between them is not hopelessly high, especially if the transition is seen in limited simulation time. In early sampling, thus, establishing a model’s adjacency matrix is an objective goal for capturing the model’s qualitative aspects. We envision that an adjacency-based scheme might be used initially to establish the model’s connectivity, after which eigenvalue-based sampling could be used to refine the quantitative nature of the state-to-state transition probabilities.

In this Article, we first offer a description of the general topology of protein folding MSMs based on the well-known graph structure of the Cayley tree. With this knowledge, we proceed to design and test sampling schemes under a metric of adjacency-based sampling, and we report the most promising candidates for early sampling refinement.

METHODOLOGY

For our analysis of MSM topologies, mean first passage time distributions (from all states to a particular state) are used to illustrate a model’s kinetic properties. To calculate mean first passage times (MFPTs) efficiently, we employ the formalism of the fundamental matrix for ergodic Markov chains. Given a transition matrix for an ergodic aperiodic Markov chain, the fundamental matrix \mathbf{Z} is given by the formula

$$\mathbf{Z} = (\mathbf{I} - (\mathbf{T} - \mathbf{W}))^{-1}$$

where \mathbf{I} is the identity matrix, \mathbf{T} is the chain’s transition matrix, and \mathbf{W} is the limiting matrix of the transition matrix.¹⁵ The mean first passage time from a state i to a state j , m_{ij} , is then simply given by

$$m_{ij} = \frac{z_{ji} - z_{ij}}{\pi_j}$$

where π represents the stationary distribution of the chain.¹⁵

To test the effectiveness of various sampling schemes for adjacency-based sampling, we’ve elected to run Markov chain Monte Carlo (MCMC) trajectories a posteriori on toy model transition matrices and previously generated MSM transition matrices for Fs peptide, the WW domain, and the villin headpiece domain. Trajectories are truncated at 10 state-to-state transitions to simulate the short runs typical in MD simulations performed with distributed computing. Transitions occur or fail to occur based on the Metropolis acceptance criterion, with acceptance probabilities defined by the transition probabilities of the original model.¹⁶

Individual trajectory data are collected into transition count matrices: if two states i and j are adjacent to one another in a trajectory vector, a count of “1” is placed in the (i,j) th entry of a matrix of dimension $N \times N$, where N is the number of states in the predefined model. After a set number of individual trajectories have run to completion, the aggregate count matrix can be normalized to yield a transition matrix, which can be compared to the “exact” matrix of the model.

In evaluating success in adjacency-based sampling, the adjacency error, or the number of missed connections in the sampling-generated matrix, serves as a reasonable metric.

Formally, the adjacency error ($\sigma_{\text{adjacency}}$) is given by

$$\sigma_{\text{adjacency}} = \sum_{ij} (A(\mathbf{T}) - A(\mathbf{T}^*))$$

where $A(\mathbf{T})$ is the adjacency matrix of the transition matrix for the predefined model and $A(\mathbf{T}^*)$ is the adjacency matrix for the sampling-generated matrix. It should be noted that, in this scheme, it is impossible for the sampling-generated matrix to have connections that are absent in the original matrix.

One round of sampling consists of evaluating count matrix rows based on a certain criterion (e.g., fewest counts or greatest contribution to eigenvalue uncertainty) and starting a new trajectory based on the results of the analysis. Eigenvalue-based sampling code was based on that presented in the literature, wherein simulations are started from the state that contributes most to uncertainty in the model’s slowest rate (largest nonunit eigenvalue).^{12,13} Even sampling distributes trajectories uniformly among already discovered states; count-based sampling favors previously discovered states with the fewest aggregate counts (i.e., the states that have been visited the fewest number of times in the simulation). Finally, connectivity-based sampling starts trajectories from the already discovered state, which is least connected (the state with the fewest adjacency matrix entries). Values of adjacency errors reported correspond to averages over 100 simulations run.

Toy models with inward direction were prepared by setting “inward” transition probabilities at greater values than “outward” probabilities. For the Cayley tree, one vertex was designated the root of the tree, and trajectories were directed toward the root. Similarly, one vertex of the hypercube model was designated as a sink, and transitions to vertices more proximal to that sink were favored with higher probabilities. Specifically, inward-directed edges were weighted so that an inward transition occurred with 2/3 probability. Diffusive models were constructed so that all transitions between adjacent nodes occurred with equal probability.

RESULTS AND DISCUSSION

Topological Characterization. In a recent publication, Bowman and Pande use mean first passage time distributions to illustrate the native state’s role as a kinetic hub: mean first passage times to the native state were observed to be shorter than those to unfolded states, suggesting that the native state serves as a hub between unfolded states.⁸ The authors also note that no unfolded states are more than two connections separated from the native state and classify states as “unfolded” (not directly connected to the native state) and “intermediate” (directly connected to the native state). Figure 1c shows a high-resolution histogram of the mean first passage time distribution to the native state for the villin macrostate MSM. While the histogram is noisy, two prominent peaks are clearly present in the plot. Corroborated by direct inspection of calculated MFPTs, the proximal peak indeed corresponds to the intermediate states and the distal to the unfolded states.

One feature of MFPTs to unfolded states is also notable: MFPTs to an unfolded state are sharply distributed around the mean first passage time from the native state to that unfolded state. Table 1 contains selected data to illustrate this relationship.

Considering these two observations, we can draw some general conclusions about dynamics on the villin macrostate network. Importantly, trajectories appear to reach the native state in a rapid enough manner to discriminate generational origin, that is, from

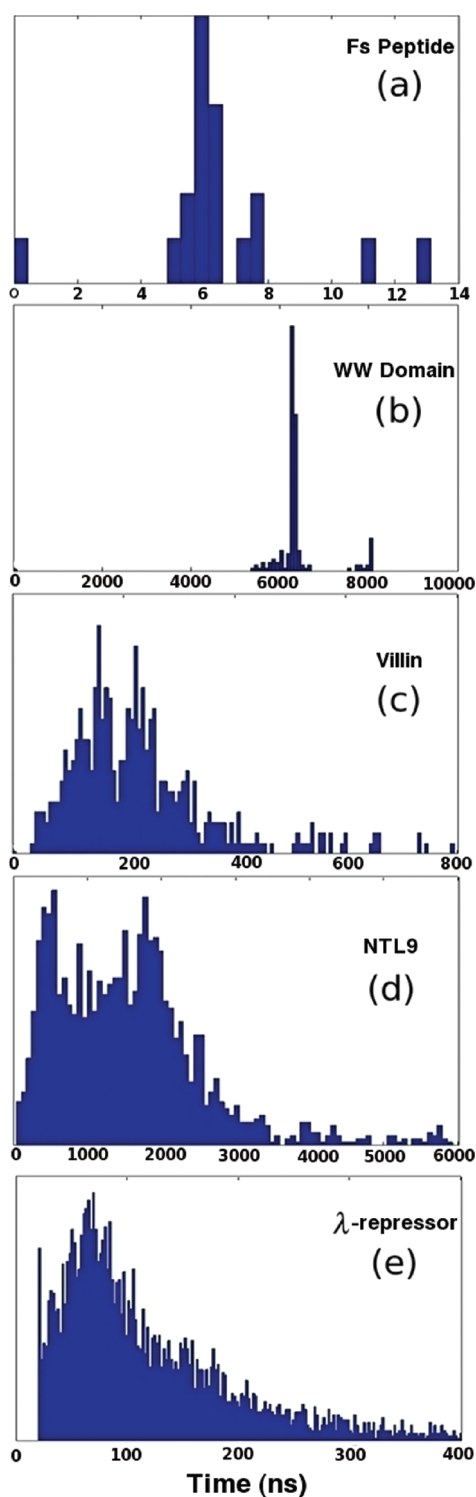


Figure 1. Mean first passage time distributions from the unfolded to native states of various protein folding MSMs: (a) Fs peptide, at 19 states with lag time 2 ns, (b) WW domain, at 200 states with lag time 35 ns, (c) villin headpiece domain, with 500 states at lag time 10 ns (d) NTL9, with 2000 states at lag time 20 ns, and (e) λ -repressor four-helix bundle with 5000 states at lag time 20 ns.

either the unfolded or intermediate states. By contrast, MFPTs to the majority of unfolded states seem to be independent of origin, suggesting that moving from native to unfolded is the rate-limiting

Table 1. MFPT Distributions among Unfolded States, Villin Headpiece MSM

particular unfolded state, U^*	center of MFPT distribution, unfolded states to U^*	MFPT, native state to U^*
5	15 908	15 905
126	9212	9207
350	3885	3881

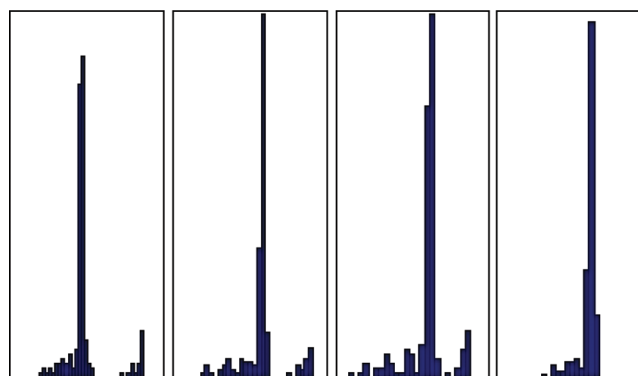


Figure 2. Noise progression in MFPT distribution from unfolded states to native state in the WW domain. The noise floor on the WW domain transition matrix was systematically raised through addition of noise from the first panel to the last, and related MFPTs were calculated for the new transition matrix. From left to right, histograms represent distributions for transition matrices with Gaussian noise ($\langle x \rangle \approx 0.1$, $\sigma_x \approx 0.05$) added to 0% of states, 0.1% of states, 1% of states, and 10% of states.

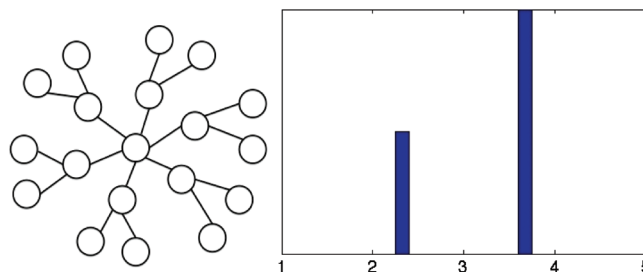


Figure 3. Left: General graph structure of an irregular Cayley tree truncated after two generations. Right: MFPT distribution from “leaf” states to “root” state in inward-directed dynamics. In this case, inward dynamics are defined such that the probability of an inward transition is $2/3$.

step in passage between most unfolded states. We can thus conclude that dynamics “inward” toward the native state are fundamentally fast, while those “outward” toward most other unfolded states are typically slow.

Figure 1a,b and d,e shows high-resolution histograms of MFPT distributions to the native states of four other protein folding macrostate MSMs. First, we should note again that very few states in any model are more than two connections removed from the native state. Second, as with villin, two peaks are readily evident in each of the first three distributions, suggesting the same generational behavior seen in villin is present in models of Fs peptide, the WW domain, and NTL9. The lack of two distinct peaks in the λ -repressor distribution, we assert, can be partially attributed to noise due to sampling limitations.

Table 2. Adjacency Error Rankings for Hybrid Sampling Schemes

method	inward Cayley tree	inward hypercube	diffusive Cayley tree	diffusive hypercube	random matrix	Fs peptide	WW domain	villin
eigenvalue-based sampling	4	4	3	4	2	4	4	4
count-based sampling	1	1	1	1	1	1	1	1
connectivity-based sampling	2	2	4	2	4	3	3	3
even sampling	3	3	2	3	3	2	2	2

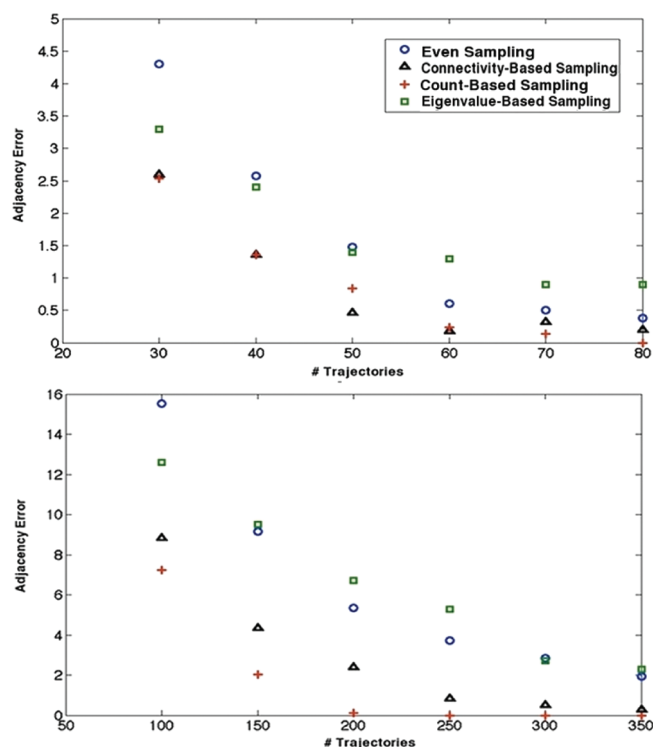


Figure 4. Adjacency error as a function of trajectory number for an inward directed irregular Cayley tree (top, 19 states) and an inward-directed five-dimensional hypercube (bottom). In both cases, count-based sampling seems to perform the best among all methods tested.

As justification for this claim, observe that “noisiness” in distributions is directly correlated with model and system size (see Figure 1, caption). To illustrate the effects of sampling error noise on MFPT distributions, Figure 2 shows an added-noise progression of the WW domain unfolded to native state MFPT distribution.¹⁷ Clearly, the two distinct peaks in the MFPT distribution merge into one broader peak as more random error is added. We postulate that, with more exhaustive sampling, two peaks would also become evident in the λ -repressor model. The rate-limiting nature of the native state in passing from unfolded to unfolded state was observed in all four models.

One simple graphical model shares many of the properties demonstrated by our MSMs: the n -irregular rooted tree, commonly known as the Cayley tree. If we truncate the Cayley tree after two generations (expanding the first generation to match the number of intermediate states in a typical MSM) and direct the dynamics in toward the tree’s root, we indeed observe kinetic behavior similar to that of protein folding MSMs. It should be noted that in protein folding MSMs, individual intermediate states are connected to relatively few (i.e., 2 or 3) unfolded states, justifying the first generation expansion of the Cayley tree.

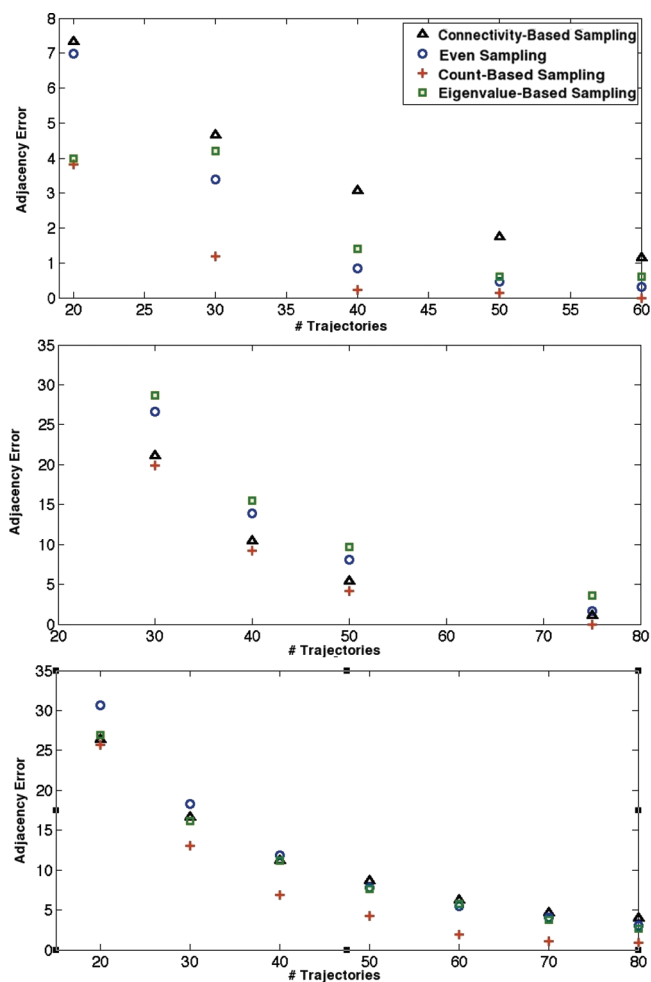


Figure 5. Adjacency error as a function of trajectory number for a diffusive Cayley tree (top, 19 states), a diffusive five-dimensional hypercube (middle), and a randomly connected matrix of density 1/2 (bottom, 19 states).

Figure 3 provides an illustration of a small irregular Cayley tree and shows the mean first passage time distribution to the representative tree’s root. MFPTs to the tree’s root are generational in nature, and pathways between tips of leaves are rate-limited by passage outward from the native state. We thus suggest that the irregular, inward-directed Cayley tree serves as an (albeit simplified) framework for thinking about protein folding MSM graphical topologies.

Adjacency-Based Sampling. One area in which this general topological characterization promises to be useful is that of sampling scheme design. Assuming protein folding MSMs have the general kinetic characteristics of inward-directed Cayley trees, we know that connectivity can best be explored by starting simulations from states far from an MSM’s “root”. While eigenvalue-based sampling may

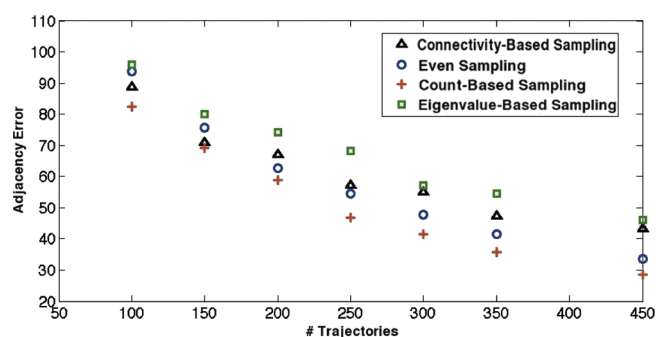


Figure 6. Adjacency error as a function of trajectory number for the Fs peptide transition matrix.

indirectly target these states, more direct methods to optimize topological exploration can certainly be conceived. In particular, we introduce two new sampling schemes called “count-based sampling” and “connectivity-based sampling”. In count-based sampling, new simulations are started from states with the fewest counts; connectivity-based sampling favors those states with the fewest connections to other states (i.e., the fewest entries in the adjacency matrix). As distant states in a Cayley tree are the least visited and least connected, we hypothesize that these methods will be effective in rapidly exploring adjacency in our models.

To test this hypothesis, we have run a series of a posteriori MCMC trajectories on predefined MSM transition matrices and have computed average adjacency errors for the generated count matrix. Adjacency-error rankings (with ranking “1” corresponding to the smallest error, based on the final column of data points in each error plot) for all methods over all models are summarized in Table 2.

Figures 4 and 5 contain plots of adjacency error versus number of trajectories for three toy models used in this study: the random stochastic matrix, the n -dimensional hypercube, and the irregular Cayley tree. In the case of the latter two models, both diffusive and inward-directed dynamics were tested under the various sampling schemes.

Although the relative performance of each method varied from toy model to toy model, two consistencies in the data are glaring: (1) that count-based sampling performs best in discovering graph adjacency and (2) eigenvalue-based sampling often performs worst at the same task. Connectivity-based sampling is successful on the hypercube, where connectivity is regular and extensive, but is relatively poor at capturing adjacency elsewhere as compared to the count-based method. The magnitudes of entries in the count matrix, therefore, seem to play an important role in defining states around which topology is poorly explored. These preliminary results suggest count-based sampling would provide effective adjacency determination early in model construction.

The next test of sampling effectiveness involves sampling on pre-existing MSM transition matrices. Figure 6 shows adjacency error versus number of trajectories for the four sampling schemes on the Fs peptide transition matrix. In agreement with the toy model analyses, count-based sampling performed the best among all schemes, while eigenvalue-based sampling behaved the worst. In this case, even sampling proved a better technique than connectivity-based sampling, underlining the apparent importance of count magnitudes for adjacency-based sampling. Figure 7 contains plots of adjacency error versus number of trajectories for the WW domain and the villin headpiece domain. Both eigenvalue-based

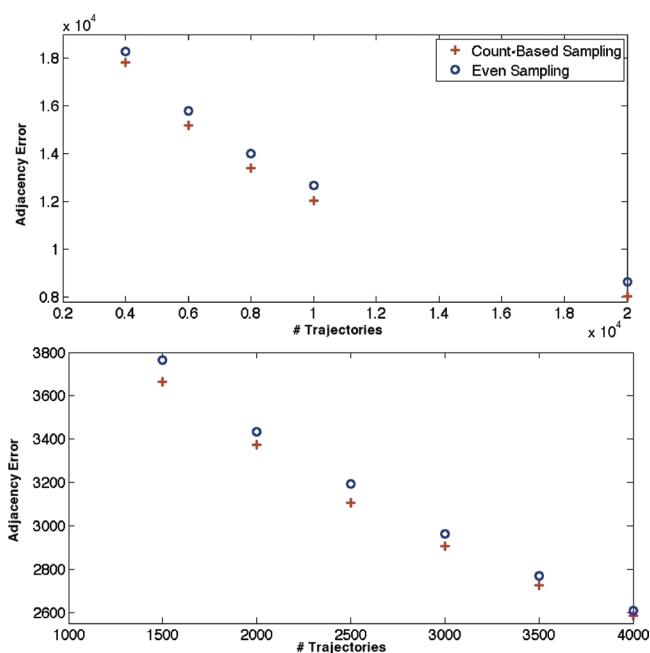


Figure 7. Adjacency error as a function of trajectory number for the WW domain (top) and villin headpiece domain (bottom) transition matrices. Connectivity-based sampling and eigenvalue-based sampling, which would appear well above the two methods shown in adjacency error, are omitted to preserve scale.

sampling and connectivity-based sampling performed radically worse than the other two schemes; only even and count-based sampling are shown in the figure to preserve scale. Clearly, count-based sampling performs better than even sampling for both systems. For all of the above systems, we thus conclude that count-based sampling is the best strategy for capturing adjacency among those tested.

While adjacency provides some description of MSM dynamics, quantitative transition probabilities are obviously important in building meaningful models. To evaluate effects of sampling on absolute transition matrix error, we introduce a hybrid sampling scheme, which combines explorative and eigenvalue-based sampling methods. Explorative sampling serves to solidify state definitions at early stages in model building; once states are well-defined (i.e., discovered), eigenvalue-based sampling should refine values for transition probabilities as intended. One might employ an adjacency error cutoff to determine the point at which switching from one sampling method to the other would be appropriate.

Figure 8 shows the absolute error in the Fs peptide transition matrix generated by four different hybrid sampling schemes. To facilitate comparison among methods, each variable type of sampling (e.g., even, count-based, or connectivity-based) is carried out for 1000 trajectories and followed by the requisite number of eigenvalue-based trajectories to reach an absolute error cutoff of 2.00. For the case of pure eigenvalue-based sampling, trajectories generated only from that sampling method were used to build the transition matrix.

As is clear from the figure, count-based sampling in conjunction with eigenvalue-based sampling converged most quickly to the error tolerance for the exact transition matrix. Purely eigenvalue-based sampling, by contrast, converged more slowly than any of the hybrid sampling schemes, and took more than an order

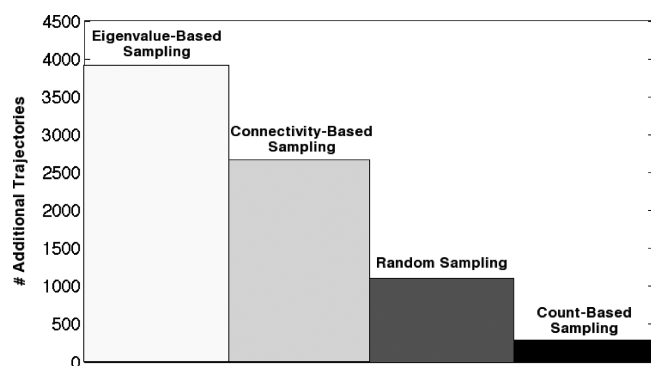


Figure 8. Convergence time for Monte Carlo Fs peptide transition matrix generated with various hybrid sampling schemes. Time is measured in the number of eigenvalue-based trajectories needed to converge to an absolute error of 2.00 after 1000 initial trajectories are run from a chosen sampling method. Absolute error is defined as the sum of absolute deviations in transition matrix elements. Convergence times for each method were, averaged over 10 simulations, (1) 3913 for pure eigenvalue-based sampling, (2) 2669 for connectivity-based hybrid sampling, (3) 1107 for even sampling hybrid sampling, and (4) 286 for count-based hybrid sampling.

of magnitude more trajectories to converge than did the best method tested. We attribute this difference, again, to an adjacency error effect: while count-based sampling had discovered >95% of states in 1000 trajectories, eigenvalue-based sampling lagged far behind at approximately 80%. These results suggest that in a finite sampling period, some type of hybrid sampling will perform better than eigenvalue-based sampling, and among the hybrid sampling techniques tested, count-based sampling is the most effective.

CONCLUSION

We thus observe that protein folding MSMs have certain general topological characteristics, and we see that these characteristics can be used to design directed sampling schemes for MSM construction. The next step in evaluating the hybrid sampling schemes discussed above would entail actually testing them on systems at the molecular dynamics (MD) level. We suggest that such hybrid sampling schemes could easily be tested directly or in an environment in which MSMs drive sampling (e.g., Copernicus), wherein plug-in modules for different sampling techniques could be swapped in and out without effort.¹¹ Given the high degree of sampling already performed on molecules like villin and various helical peptides, such systems could serve as ideal candidates against which various sampling schemes could be benchmarked.

Direct application of these hybrid sampling methods to MD simulations would be straightforward in concept, comprising the iteration of three steps: (1) running a series of short MD trajectories, (2) building an MSM based on the aggregate data, and (3) seeding new MD trajectories based on the sampling criterion (e.g., from the states with the fewest counts). Effective use of hybrid sampling techniques in MD studies could allow for the generation of accurate MSMs from a minimal set of short trajectories, enhancing both model accuracy and sampling efficiency.

While these hybrid sampling methods will be easily extensible to MD simulations, one will need to use MSM error metrics alternative to those used with MCMC in this Article. MSMs constructed from MD data are generated in a partially stochastic

fashion, making error evaluation based on numerical properties of the transition matrix impractical. MSM observables (like native state identity and stability, eigenspectral properties, and projections onto experimental observables) and uncertainties therein will instead need to form the basis for comparison between models and validation of hybrid sampling methods. We do envision, however, that convergence of model size could serve as an adjacency-like metric: when new states cease to appear after iteration of the above procedure, the sampling scheme could be changed to the eigenvalue method. Final eigenvalue-based sampling could then be carried out to a satisfactory threshold defined by model observable uncertainties.

As a second caveat, we should note that at high temperature (i.e., well above biological temperatures), protein folding networks become more connected and thus lose some degree of the tree-like structure identified in this study. Accordingly, one should take care in using the sampling algorithms developed here in high temperature simulations. However, as count-based sampling performed well even on a randomly connected graph (see Figure 5), we expect our hybrid algorithms to remain effective in systems held at higher temperatures.

We acknowledge that the exploration in the sense of a posteriori sampling in this Article is somewhat contrived. After all, only states that exist on the underlying MSM network can ever be discovered. Provided enough time, even eigenvalue-based sampling would capture all of the adjacency of a transition matrix after the model has been completely constructed.

However, given that count-based sampling discovers states in a much more computationally efficient fashion than eigenvalue-based sampling, we posit that improved performance due to count-based sampling will translate to the arena of atomistic simulations. After all, we assume that a predefined network underlies all dynamics in atomistic simulation: the network of the system's free energy landscape. It is the nature of this network that we seek to explore in performing molecular dynamics simulations.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

ACKNOWLEDGMENT

We thank T. J. Lane, Sergio Bacallado, Greg Bowman, and Kyle Beauchamp for providing data and insight to facilitate this study. We thank the NSF (MCB-0954714) and the NIH (R01-GM062868) for their support of this work. J.K.W. was supported by the Fannie and John Hertz Foundation on the endowed Professor Yaser S. Abu-Mostafa Fellowship.

REFERENCES

- (1) Noé, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **2008**, *128*, 244103.
- (2) Berezhovskii, A.; Hummer, G.; Szabo, A. Reactive flux and folding pathways in network models of course-grained protein dynamics. *J. Chem. Phys.* **2009**, *130*, 205102.
- (3) Yang, S.; Banavali, N. K.; Roux, B. Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 3776–3781.
- (4) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.

(5) Schutte, C. Conformational dynamics: modeling, theory, algorithm, and application to biomolecules. Habilitation thesis, Department of Mathematics and Computer Science, Freie Universitat Berlin, 1999.

(6) Swope, W. C.; Pitera, J. W.; Suits, F. Describing protein folding kinetics by molecular dynamics simulations. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(7) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.

(8) Bowman, G. R.; Voelz, V. A.; Pande, V. S. Atomistic folding simulations of the five-helix bundle protein λ_{6-85} . *J. Am. Chem. Soc.* **2011**, *133*, 664–667.

(9) Bowman, G. R.; Pande, V. S. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890–10895.

(10) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009**, *131*, 124101.

(11) Pronk, S.; Larsson, P.; Pouya, I.; Bowman, G. R.; Haque, I.; Beauchamp, K.; Hess, B.; Pande, V. S.; Kasson, P.; Lindahl, E. *Copernicus: A New Paradigm for Parallel Adaptive Molecular Dynamics*; KTH Royal Institute of Technology: Stockholm, Sweden; Stanford University: Stanford, CA; University of Virginia: Charlottesville, VA. Unpublished work, 2010.

(12) Singhal, N.; Pande, V. S. Error analysis in Markovian State Models for protein folding. *J. Chem. Phys.* **2005**, *123*, 204909.

(13) Hinrichs, N. S. Algorithms for building models of molecular motion from simulations. Ph.D. Dissertation, Stanford University, 2007.

(14) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. Adaptive seeding method: rapid equilibrium sampling initiated from non-equilibrium data. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19765–19769.

(15) Peskun, P. H. Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **1973**, *60*, 607.

(16) Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611–6615.

(17) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories*; Stanford University: Stanford, CA. Unpublished work, 2011.

MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale

Kyle A. Beauchamp,[†] Gregory R. Bowman,^{‡,⊥} Thomas J. Lane,[‡] Lutz Maibaum,[‡] Imran S. Haque,[§] and Vijay S. Pande^{*,†,‡,§}

[†]Biophysics Program, [‡]Chemistry Department, [§]Computer Science Department, Stanford University, Stanford, California

ABSTRACT: Markov state models provide a framework for understanding the fundamental states and rates in the conformational dynamics of biomolecules. We describe an improved protocol for constructing Markov state models from molecular dynamics simulations. The new protocol includes advances in clustering, data preparation, and model estimation; these improvements lead to significant increases in model accuracy, as assessed by the ability to recapitulate equilibrium and kinetic properties of reference systems. A high-performance implementation of this protocol, provided in MSMBuilder2, is validated on dynamics ranging from picoseconds to milliseconds.

1. INTRODUCTION

Conformational changes such as myosin procession,¹ protein folding,² and ligand binding³ have long occupied the attention of biophysicists. A predictive, first-principles understanding of conformational dynamics could elucidate these processes in atomic detail, with broad applications in engineering and medicine. Many biophysical experiments probe the fundamental states and rates of a system. For example, the dominant conformational state of a biomolecule can be determined experimentally by NMR spectroscopy⁴ or X-ray crystallography,⁵ while the existence of intermediate states can be demonstrated by kinetic studies.^{6,7} Even at the single-molecule level, dynamics between multiple conformational states can be tracked by monitoring observables (e.g., FRET)⁸ that report on the conformational details of a molecule. Conformational states and their rates of interconversion remain a unifying paradigm of biophysical studies.

Discrete-time master equations, or Markov state models,^{9–11} formalize this paradigm. In a Markov state model, one defines a set of conformational states and models the dynamics between them as a Markov jump process on that state space. Predicted conformational states and rates can be extracted from atomistic molecular dynamics simulations of biomolecular dynamics under ambient conditions.^{12–14} Here, we describe an improved protocol for constructing Markov state models from an ensemble of molecular dynamics simulations. This enhanced protocol has been implemented as version 2.0 of the freely available MSMBuilder software package, available at <https://simtk.org/home/msmbuilder>. The improvements in MSMBuilder2 include a more accurate state definition through hybrid k-centers k-medoids clustering, improved estimates of kinetic and equilibrium properties via a reversible maximum likelihood estimator,^{9,11} and an extensible Python implementation allowing facile customization. We validate and benchmark the protocol on proteins spanning a range of time scales and sizes.

2. THEORY

A Markov state model^{9,10,15–17} consists of a set of state definitions and a transition probability matrix characterizing the kinetics on this state space. In this work, we adopt the

following conventions. States are labeled integers $\{1, 2, \dots, n\}$. Transition matrix entry ij gives the conditional probability of jumping from state i to state j during a time interval (lagtime) τ :

$$T_{ij}(\tau) = P(\sigma(x(\tau)) = j | \sigma(x(0)) = i) \quad (1)$$

where $\sigma(x)$ is a function mapping the conformation x onto the state space. Equilibrium conformational dynamics are expected to satisfy detailed balance: that is, $\pi_i T_{ij} = \pi_j T_{ji}$ where π_i is the equilibrium population of state i . Because of the symmetry of the detailed balance equation, we define a symmetric matrix $X_{ij} = X_{ji} = \pi_i T_{ij}$. This matrix gives the counts between states i and j at equilibrium, normalized such that $\sum_j X_{ij} = 1$. With this definition, the transition matrix can be expressed as $T = D^{-1}X$, where $D = \text{diag}(\pi)$ is a diagonal matrix of equilibrium populations.

The eigenvalues and eigenvectors of a transition matrix have special significance. Let (λ_i, v_i) be an eigenvalue–eigenvector pair for T (e.g., $Tv_i = \lambda_i v_i$). By comparison to the eigenvalues $(1/\tau_i)$ of a continuous-time master equation rate matrix K , one can show that the eigenvalues of a transition matrix are related to the relaxation time scales (τ_i) of a master equation via $\lambda_i = \exp(-\tau/\tau_i)$, where τ is the lagtime used to estimate the transition matrix.^{15,18} For systems satisfying detailed balance, the eigenvalues λ_i must be real, as the eigenvalue equation can be written as a symmetric generalized eigenproblem: $Xv_i = \lambda_i Dv_i$. We point out that a recent work⁹ provides an excellent review of the theory of MSMs; another review covers both theoretical and experimental aspects as applied to protein folding.¹⁹

To estimate a transition matrix, one must fix a lagtime, which we signify by writing transition matrices with explicit lagtime dependence $T(\tau)$. Because they describe physical observables, relaxation time scales should be insensitive to changes in lagtime. However, projecting dynamics onto a finite state space results in dynamics that are only approximately Markovian. Thus, a common test of model consistency is to calculate the relaxation time scales for a sequence of lagtimes.^{9,10,18} In practice, discretization error manifests itself as erroneously fast time scales for

Received: July 5, 2011

Published: September 06, 2011

short lagtimes. Indeed, it has been shown^{9,20} that increasing either the number of states or the lagtime will lead to more accurate models; however, finite sampling and computational resources place limits on the number of states and lagtime.

3. METHODS

This paper presents the recent advances in MSMBuilder2. Below, we discuss these advances, in terms of both the nature of the improvement as well as its motivation. We propose the following new protocol for MSM construction, which shares some characteristics with ones previously developed by ourselves and others.^{9,11,21}

1. Cluster molecular dynamics trajectories using a hybrid k-centers k-medoids algorithm.
2. Restrict data to its maximal ergodic subgraph.
3. Estimate transition and count matrices ($T(\tau)$, $C(\tau)$) using a maximum likelihood reversible estimator.

While this protocol is similar to previous approaches in broad strokes, these key refinements make the approach more quantitative without increasing computational cost. We note that MSMBuilder2 also allows nonreversible maximum likelihood estimation for systems where reversibility is not desired.

3.1. Hybrid k-Centers k-Medoids Clustering. The first step in MSM construction is to identify conformational states. Because MSM accuracy depends on the quality of state decomposition, enhanced clustering is a natural way to improve MSM methods. In MSMBuilder2, as in other MSM methods, it is vital to achieve *kinetic* clustering—that is, states sufficiently fine so as to be free from internal kinetic barriers.

Previous work^{9,11} used an $O(kN)$ approximate k-centers clustering,²² where k denotes the desired number of clusters and N denotes the number of conformations. That algorithm can be viewed as an approximate solution to the problem:

$$\min_{\sigma} \max_i d(x_i, \sigma(x_i)) \quad (2)$$

Here, $\sigma(x)$ is the “assignment” function that maps a conformation to the nearest cluster center. $d(x,y)$ is the distance between two conformations x and y , measured via the RMSD metric.²³ The minimization occurs over all clusterings (σ) with k states, subject to some choice of initial center. Finally, the max is taken over all conformations in the data set.

The k-centers approach minimizes the worst-case clustering error, as quantified by the objective function $f_{\max}(\sigma) = \max_i d(x_i, \sigma(x_i))$. Considering only the worst-case clustering error is problematic for conformational dynamics, particularly in protein folding, as the worst-case error is often determined by extended (unfolded) conformations with very small populations. Furthermore, cluster centers generated by this algorithm are often noncentral; that is, they often do not represent the geometric center of their associated data.

Alternatively, k-medoids clustering²⁴ approximately minimizes $f_{\text{med}}(\sigma) = (1/N) \sum_i d(x_i, \sigma(x_i))^2$. With sufficient sampling, constant temperature molecular dynamics draws Boltzmann-weighted conformations; thus, by averaging over all conformations, $f_{\text{med}}(\sigma)$ is an objective function that penalizes the (approximately) ensemble-averaged deviation from cluster centers. The resulting clusters tend to be centrally located within their respective data—i.e., they are medoids.²⁵ However, for folded proteins, strict Boltzmann weighting yields few unfolded states, often leaving unfolded conformations assigned to folded

states. This deficiency can be explained in terms of $f_{\max}(\sigma)$. A clustering that minimizes $f_{\text{med}}(\sigma)$ may in fact be worse when evaluated by $f_{\max}(\sigma)$; conversely, minimizing $f_{\max}(\sigma)$ could increase $f_{\text{med}}(\sigma)$. For accurate kinetic clustering of biomolecule dynamics, one should consider *both* the worst case (f_{\max}) and average case (f_{med}) clustering error.

Simultaneously optimizing both the average and worst-case error can be achieved by combining the k-centers and k-medoid algorithms. Let ε be some desired worst-case clustering error. Define the set

$$S(\varepsilon) = \{\sigma : f_{\max}(\sigma) \leq \varepsilon\} \quad (3)$$

Thus, $S(\varepsilon)$ is the set of all clusterings that have worst-case errors of ε (or better). We now apply a k-medoids clustering algorithm, but restricted to the set $S(\varepsilon)$. In practice, we use a two step approach:

1. Apply approximate k-centers to return initial clusters g_i , terminating when $f_{\max}(\sigma) \leq \varepsilon$.
2. Apply approximate k-medoids to the result, but rejecting all moves that increase $f_{\max}(\sigma)$.

For step 2, we employ a modification of the partitioning across medoids algorithm.²⁴ For each cluster g_i , we randomly select a conformation x_i assigned to that state. The clustering errors (f_{med}, f_{\max}) are calculated and compared to the values that would be obtained were x_i instead the cluster center of that state. If f_{med} is improved and f_{\max} is improved (or unchanged), the move is accepted. In practice, f_{\max} decreases insignificantly during this process, but f_{med} decreases dramatically over a handful of iterations. As described, the hybrid algorithm tends to preserve the overall distribution of clusters, essentially refining k-centers to be more “central”; this is desirable because k-centers is known²² to provide a reasonable partition of conformation space.

3.2. Improved Estimators for Reversible Transition and Count Matrices. Since equilibrium conformational dynamics obeys detailed balance, it is important for MSMs to satisfy detailed balance (also called reversibility). A positive reversible MSM guarantees positive real eigenvalues λ , which can be interpreted as relaxation time scales through the relation $\tau_{\text{rel}} = -\tau_{\text{lag}}/\log(\lambda)$. Previous work¹¹ has used the symmetrized counts—so-called because the count matrix is symmetrized via the equation $C' = 1/2(C + C^T)$ —to estimate a reversible count matrix. Though the resulting MSMs satisfy detailed balance, this estimator can introduce artifacts in both equilibrium and kinetic properties;^{15,21} this error is pronounced for short trajectories started from a distribution far from the system’s equilibrium. A recent work²¹ recommends estimating a transition matrix using the unsymmetrized counts after restricting the data to its maximal ergodic subgraph. Thus, after clustering, one must first identify the maximal ergodic (i.e., strongly connected) subgraph—that is, a (maximal) set of states M such that if $i \in M$ and $j \in M$, then there exists a path from $i \rightarrow j$ and from $j \rightarrow i$. That approach eliminates artifacts in equilibrium estimates but yields transition matrices that may not satisfy detailed balance. To enforce detailed balance while preserving accurate estimation of equilibrium properties, we have implemented the following protocol:

1. Apply Tarjan’s algorithm,²⁶ restricting data to the maximal ergodic subgraph.
2. Estimate a reversible count matrix using a maximum likelihood estimator.

The theory of reversible estimation has been discussed previously;^{9,11,16,27,28} however, several implementation issues

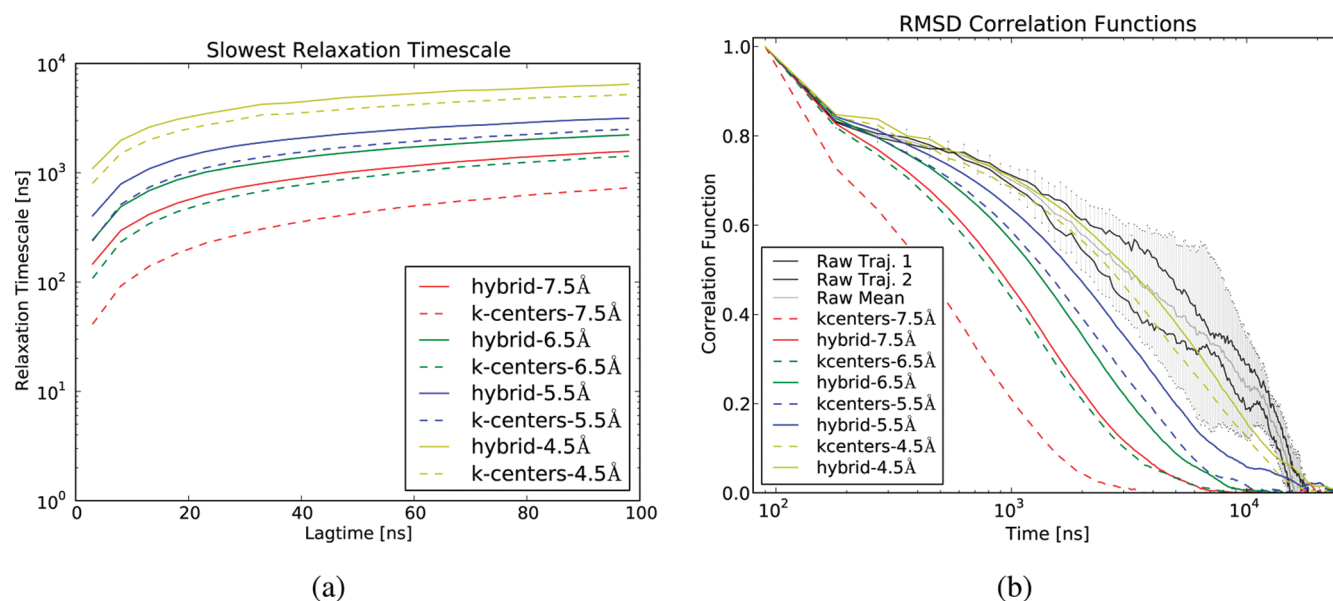


Figure 1. (a) Relaxation time scales of models constructed with k-centers and hybrid clustering. (b) RMSD correlation functions as calculated by different clusterings. MSMs in b constructed with 90 ns lagtime. MSMs constructed from simulations of the WW protein; see Appendix 1.

have limited its general use. First, the reversible MLE estimator is only well-defined for ergodic MSMs, so the trimming procedure is critical. Second, the iterative procedure sometimes converges slowly for many-state models; in Appendix 2, we discuss an efficient implementation that allows scaling to biological systems with tens of thousands of states.

4. RESULTS

We now validate the revised MSM protocol. First, we show that improved clustering results in more self-consistent models, as measured by either relaxation time scales or correlation function analysis. Second, we show that improved transition matrix estimators result in an improved ability to recapitulate kinetic and equilibrium properties of a known reference model.

4.1. Hybrid k-Centers k-Medoids Clustering Improves State Definitions. Projecting onto a finite state space results in dynamics that are only approximately Markovian. One way to evaluate model consistency is by calculating the relaxation time scales for a sequence of lagtimes; as observables, these time scales should be approximately lagtime-independent. As compared to models constructed with k-centers clustering, hybrid clustering yields relaxation time scales that are slower (Figure 1a) and less lagtime-dependent. For models with few states ($f_{\max} = 5.5\text{--}7.5$ Å; Table 1), hybrid clustering performs considerably better than k-centers. In particular, a hybrid model with a fixed number of states (e.g., 176 states, or $f_{\max} = 7.5$ Å) performs comparably with a k-centers model with considerably more states (e.g., 806 states, or $f_{\max} = 6.5$ Å). In the limit of many states, hybrid and k-centers perform comparably, as eventually both k-centers and the hybrid yield one state per sampled conformation; however, statistically accurate estimation is impossible when the number of states approaches the total number of available conformations. For this reason, it is desirable to achieve accurate models with as few states as possible.

The lack of a true reference value makes relaxation time scales an incomplete validation of MSM kinetics. Correlation function analysis offers an orthogonal check with a known reference value.

Table 1. Models Constructed from WW Domain Simulations Were Used to Compare Structural Properties of k-Centers and Hybrid Clusterings^a

model	# states	f_{\max} (Å)	f_{med} (Å)
k-centers	26104	4.5	2.97
hybrid	26104	4.5	2.21
k-centers	5135	5.50	4.21
hybrid	5135	5.50	2.97
k-centers	806	6.50	4.76
hybrid	806	6.48	3.60
k-centers	175	7.48	6.03
hybrid	175	7.47	3.97

^aThe number of states for each model was determined by k-centers convergence based on a prespecified f_{\max} ; hybrid clusterings use the same k-centers clusters and iteratively improve them by the algorithm described above.

The RMSD correlation function is given by $\gamma(t) = (\langle s(t)s(0) \rangle) / (\langle s(t)^2 \rangle)$, where $s(t) = r(t) - \langle r(t) \rangle$ and $r(t)$ is the RMSD to a reference structure, here taken to be the native conformation. For the MSM calculation, the transition matrix was used to first calculate a pseudotrajectory of 100 000 lagtimes (9 000 000 ns). For each frame in the pseudotrajectory, a RMSD value was randomly selected from the collection of RMSD values observed for that state. This approach models intrastate dynamics by the random selection of each RMSD value.

As compared to the reference (calculated from the raw data), MSMs with few states show erroneously fast kinetics (Figure 1b); hybrid clustering partially mitigates this error. With sufficiently many states (e.g., $f_{\max} \leq 4.5$), the dynamics are accurately captured by the MSM. Both raw and MSM RMSD correlation functions decay on a time scale comparable to the folding–unfolding dynamics of the protein. Further increasing the number of states is not feasible due to increased statistical uncertainty (Appendix 5). We observe similar results for alanine dipeptide (Appendix 6).

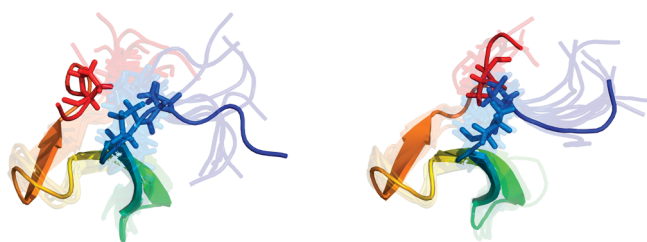


Figure 2. Cluster centers (opaque) and randomly sampled conformations (transparent) are displayed for the most populated state from models based on the k-centers and hybrid clustering algorithms. Both models are based on simulations of the WW domain. The hybrid clusters (b) were constructed by improving the initial k-centers clustering in a. Both clusterings have 806 states ($f_{\max} = 6.5 \text{ \AA}$).

In addition to enabling kinetic calculations, clustering provides an important tool for exploratory data analysis, which benefits from cluster centers that are representative of their associated data. Yet, with k-centers clustering, the f_{\max} objective function is inherently insensitive to local or average structural properties. This leads to state definitions that tend to be useful only as partitions of conformation space—in particular, minimizing f_{\max} does not ensure that cluster centers are central within their associated data. When applied to simulations of the WW protein, hybrid clustering decreases the average clustering error significantly, as quantified by the f_{med} objective function (Table 1). The hybrid clusters show less structural heterogeneity (Figure 2). Furthermore, the k-centers cluster center lacks a critical proline contact (sticks) that defines the native fold; the hybrid cluster center retains this key structural feature.

4.2. Improved Estimators for Reversible Transition and Count Matrices. The reversible MLE yields improved estimates of equilibrium and kinetic properties. As a preliminary control, the MLE and symmetrized estimators are compared on a data set consisting of two trajectories that are long ($100 \mu\text{s}$) relative to the folding and unfolding time scales ($\approx 10 \mu\text{s}$); as expected, the resulting free energies show good agreement (Figure 3).

In a more demanding test, we generate an ensemble of two-state folding trajectories from a model with a folding time scale of 100 steps and an unfolding time scale of 1000 steps (see Appendix 4). This approximates the scenario of running MD simulations from an ensemble of unfolded conformations. Because the trajectory length is comparable to the folding time scale, the symmetrized estimator biases results toward the starting distribution of conformations, which in this case is entirely unfolded.

Using the model data, transition and count matrices were estimated using the MLE and symmetrized procedures (Figure 4). The reversible MLE accurately estimates the kinetic (a,b) and equilibrium (d) properties of the reference model. However, the symmetrized estimator shows equilibrium properties that are biased toward the unfolded state (d). Furthermore, the symmetrized unfolding time scale is erroneously high (c). This symmetrization bias reduced the accuracy of some previous MSMs, as pointed out in ref 29; reversible estimation eliminates this bias.

4.3. Improved Scaling and Performance. MSM construction relies on the clustering and analysis of vast simulation data sets. For the clustering algorithms in this work, RMSD evaluations are rate limiting; further inspection shows that RMSD is bottlenecked by a matrix multiplication involving an $m \times 3$ matrix of atomic coordinates, where m is the number of atoms in each conformation. Using an SSE3-optimized matrix multiply routine³⁰

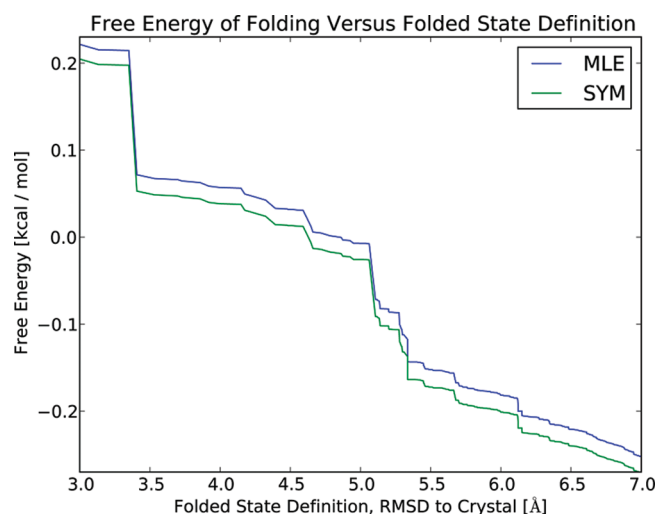


Figure 3. Simulations of the WW protein¹² were used to compare the performance of the symmetrized and MLE protocols. Folding free energies calculated using a two-state approximation ($-RT \log(\pi_{\text{folded}}/\pi_{\text{unfolded}})$), show good agreement ($\Delta \leq 0.03 \text{ kcal/mol}$) between models constructed using the symmetrized and MLE protocols, as expected for long trajectories. The near-zero folding free energy is expected, as the simulations were performed near the melting temperature;¹² the exact free energy depends weakly on how one defines the folded state. Here, the folded state is defined as all states with a RMSD (to crystal structure) below some cutoff value; the unfolded state is defined as the remaining states. The large RMSD values observed are due to the large conformational fluctuations observed in the high temperature (393 K) simulations.

with OpenMP parallelization, we have accelerated RMSD and clustering calculations by $20\times$ over the previous versions of MSMBuilder. MSMBuilder2 has been successfully applied to systems spanning a broad range of time scales and sizes; Table 2 reports the computational cost of MSM construction for various protein systems. In all cases, the cost of the MD simulations is considerably greater than the cost of MSM analysis.

5. DISCUSSION

5.1. MSMBuilder2 Protocol. As shown above, the protocol validated in this work presents several clear advantages over previous methods. These advances are evolutionary in nature, building upon previous work. The overall MSM construction protocol has retained the following key steps: perform molecular dynamics simulations, cluster data, and estimate a transition matrix. We continue to work with the RMSD metric, as its simple distance interpretation provides a physically motivated state decomposition. RMSD is a widely used distance metric for comparing biomolecular conformations;^{23,31,32} this common use allows biophysical intuition for RMSD, which is one reason for our choice of this metric. Furthermore, previous work found that, for alanine dipeptide, RMSD-based state decompositions yielded models that paralleled ones based on manual state decompositions.¹⁰ We note that some systems may benefit from other metrics; the MSMBuilder2 framework is extensible to such situations.

The procedure of kinetic clustering, whereby one leverages fine structural clustering to produce states free from kinetic barriers,^{9,10,15} benefits from the improved clustering algorithm. In kinetic clustering, it is critical to validate state decompositions

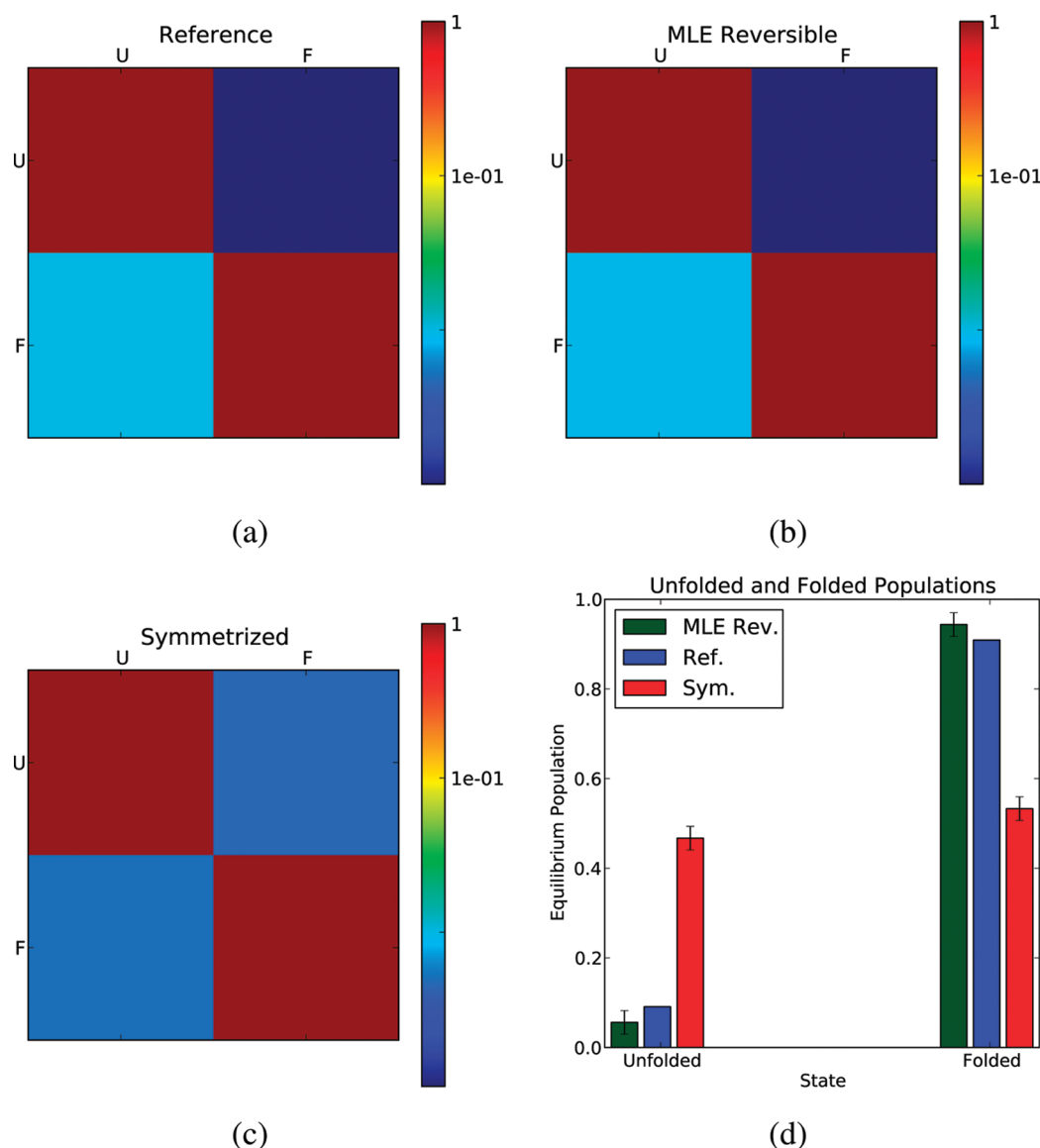


Figure 4. Simulated two-state folding simulations generated from a reference transition matrix (a) were used to estimate transition matrices. The MLE reversible procedure (b) shows good agreement with the reference transition matrix, while the symmetrized procedure (c) shows poor agreement with the reference. Furthermore, as compared to the symmetrized estimate, the MLE estimate better recapitulates the reference equilibrium properties (d).

Table 2. Computational Cost of MSM Construction for Various Protein Systems^a

system	# atoms	# frames	walltime	cluster size (f_{\max})	n_{states}	τ_{lag}	τ_{slow}
ALA	22	250000	1.0 m	0.35 Å	82	10 ps	202 ps
WW	562	200000	11.6 h	5.50 Å	26104	90 ns	5.9 μ s
HP35 (300 K)	576	109674	2.25 h	4.00 Å	9328	10 ns	7.6 μ s
λ	1258	700133	1.80 d	4.00 Å	20599	20 ns	2.0 ms

^aMSMBuilder2 was applied to various protein systems, ranging from alanine dipeptide to the λ -repressor protein. Walltimes include the cost of reading all conformations into memory, applying k-centers until convergence, and applying 10 iterations of hybrid k-medoids. The number of states is determined by applying k-centers clustering until the desired maximum cluster size f_{\max} is achieved; the hybrid step typically produces little change in f_{\max} . The slowest observed relaxation τ_{slow} is calculated by $-\tau_{\text{lag}}/\log(\lambda)$, where λ is the largest nonstationary eigenvalue of the model. τ_{lag} gives a lower bound on the time scales accessible to a given model; τ_{slow} gives an upper bound on the time scales observed in a given data set. These data suggest that the present methods can successfully model conformational dynamics from the picosecond to millisecond time scales.

using kinetic metrics; here, we have applied tests based on both relaxation time scales and correlation functions. Another key motivation for the hybrid algorithm is performance. Hybrid

clustering achieves improved clusters with only $10\times$ worse computational cost than the simple k-centers algorithm; this cost is more than offset by the accelerated RMSD calculation.

The reversible MLE protocol builds upon previous work^{9,11,21} to build accurate reversible models. Besides enforcing reversibility, the reversible MLE has other subtle benefits. First, reversibility improves statistics; because a reversible MSM is defined by a *symmetric* matrix X_{ij} , the number of possible parameters drops from n^2 to $(n(n-1))/2$. Second, the counts matrix \mathbf{X} can be visualized to gain intuition on the connectivity properties of a system. Previously, this has typically been done using transition path theory (TPT).³³ However, TPT requires *a priori* definition of initial and final states, while visualizing the counts matrix can be done in a hypothesis-free manner.

5.2. MSMBuilder2 Implementation. MSMBuilder2 is implemented as a library using the Python³⁴ language and achieves high performance by using optimized libraries (Numpy,³⁵ Scipy, Pytables³⁶) whenever possible. The rate-limiting step in clustering, the $3 \times n$ matrix multiply, is written as a small C library with Python wrappings. This design framework allows both flexibility and performance; indeed, benchmarks³⁰ suggest that the clustering code approaches the published peak efficiency of the benchmark machines. We suspect that the MSMBuilder2 library will be a useful starting point for other researchers interested in methods development. For researchers interested in applying MSMBuilder2 to analyze their simulations, the current protocol is captured by a set of command-line scripts and a tutorial at <https://simtk.org/home/msmbuilder/>.

5.3. Future Challenges. The advances in MSMBuilder2 represent significant advantages over previous methods; however, future work will likely lead to further improvements. Clustering remains a compromise between accuracy and speed. For full protein data sets ($\geq 100\,000$ conformations), performance worse than $O(kN)$ will generally be unacceptable, but other methods may further improve the results shown here. Estimation of reversible transition matrices may benefit from a Bayesian framework,^{16,27,28} accelerating such schemes for use in biological systems remains a key challenge. In addition to incremental improvements in the current protocol, more drastic changes have also been explored. In particular, other groups have shown some success working with incomplete partitions of conformation space and continuous time (master equation) modeling.^{15,18} Finally, existing frameworks consider clustering, ergodic trimming, and model estimation as three distinct steps. However, these steps are coupled and jointly contribute to modeling uncertainty. Methods that consider model accuracy and finite sampling statistics during all stages of model construction may further reduce modeling error.

6. CONCLUSION

Although modeling conformational change at atomic resolution remains challenging, the MSMBuilder2 protocol yields significant improvements in model accuracy, structural insight, and computational performance. With system sizes ranging from 22 atoms to 1258 atoms and time scales ranging from 10 ps to 2 ms, the model systems considered here suggest that MSMBuilder2 may facilitate simulation studies of previously inaccessible biomolecular systems.

APPENDICES

1. Simulation Details. Alanine dipeptide was simulated using Gromacs 4.5.3³⁷ with the AMBER96 force field and GBSA implicit solvent. One trajectory of length 50 ns was analyzed; snapshots were stored every 200 fs.

The WW domain^{38,39} simulations were described previously;¹² the authors of that work have graciously provided the trajectories

on their Web site. Simulations were performed using the AMBER99sb-ILDN⁴⁰ force field at 395 K. For MSM construction, data were stored at every 1 ns; two trajectories of length 100 μ s were analyzed.

The HP35 data set includes more than 600 simulations (minimum length 700 ns) at 300 K. Simulations were performed using Gromacs 4.5.3 with the Amber99sb-ILDN force field and TIP3P water. Conformations were stored at 1 ns intervals. Conformations were started from more than 600 different folded and unfolded conformations.

The λ -repressor simulations have been described previously.⁴¹ More than 700 simulations of minimum length 600 ns were analyzed; conformations were stored at 1 ns intervals. Simulations were performed at 370 K, using the ff03 force field with TIP3P water.

2. Maximum Likelihood Estimator for Reversible MSMs.

Suppose one has observed a matrix of counts C_{ij} ; this is typically output from the clustering and assignment stages of model construction. To estimate a general (possibly non-reversible) transition matrix \mathbf{T} , one formulates the log-likelihood function

$$f(\mathbf{T}) = \sum_{ij} C_{ij} \log(T_{ij}) \quad (4)$$

Maximizing this likelihood (e.g., ref 9) leads to the following MLE estimator of the transition matrix:

$$T_{ij} = \frac{C_{ij}}{\sum_j C_{ij}} \quad (5)$$

Suppose one knows that the underlying data are reversible. In that case, there exists a symmetric count matrix $X_{ij} = X_{ji}$ such that

$$T_{ij} = \frac{X_{ij}}{\sum_j X_{ij}} \quad (6)$$

Inserting this equation into $f(\mathbf{T})$ yields a likelihood function for \mathbf{X} , where the row sums of \mathbf{X} are defined as $X_i = \sum_j X_{ij}$ and the row sums of \mathbf{C} are defined as $N_i = \sum_j C_{ij}$:

$$f(\mathbf{X}) = \sum_{ij} C_{ij} \log(X_{ij}) - \sum_i N_i \log(X_i) \quad (7)$$

To maximize this function, one requires the partial derivatives with respect to parameters X_{ij} , which are given by $a \neq b$

$$\frac{\partial f}{\partial x_{ab}} = \frac{C_{ab} + C_{ba}}{X_{ab}} - \frac{N_a}{X_a} - \frac{N_b}{X_b} \quad (8)$$

$$\frac{\partial f}{\partial x_{aa}} = \frac{C_{aa}}{X_{aa}} - \frac{N_a}{X_a} \quad (9)$$

Setting partial derivatives to zero:

$$X_{aa} = C_{aa} \frac{X_a}{N_a} \quad (10)$$

$$X_{ab} = (C_{ab} + C_{ba}) \left(\frac{N_a}{X_a} + \frac{N_b}{X_b} \right)^{-1} \quad (11)$$

This expression can be used in an iterative update procedure. While others⁹ have suggested an approach using the quadratic formula, we find that the current formula is effective because it can be expressed entirely as simple vector and (sparse) matrix operations. In practice, we typically see

convergence within 100 000 iterations; we terminate iteration when $|\pi^{k+1} - \pi^k| \leq 10^{-10}$.

For situations with limited data, MLE estimation may require some regularization or prior to avoid overpopulating states that are strongly metastable but have been inadequately sampled. Methods to achieve regularization are discussed in the following section.

3. Incorporating Prior Pseudocounts into the Reversible MLE. It is sometimes useful to perform estimation with some nonzero prior; in practice, this involves adding a uniform matrix of pseudocounts to the observed count matrix: $C'_{ab} = C_{ab} + \alpha$. This procedure generally destroys sparsity structure, preventing its use for large systems. Below, we show a method to maintain sparsity while incorporating prior pseudocounts.

The update equation can be expressed in terms of the observed counts C_{ab} , the observed row sums N_a , the prior pseudocount (α) added at each matrix position, and the number of states n .

$$X_{aa} = (C_{aa} + \alpha) \frac{X_a}{n\alpha + N_a} \quad (12)$$

$$X_{ab} = (2\alpha + C_{ab} + C_{ba}) \left(\frac{n\alpha + N_a}{X_a} + \frac{n\alpha + N_b}{X_b} \right)^{-1} \quad (13)$$

To simplify the computation, define two intermediate variables Q_{ab} and R_{ab} :

$$Q_{ab} = (C_{ab} + C_{ba}) \left(\frac{n\alpha + N_a}{X_a} + \frac{n\alpha + N_b}{X_b} \right)^{-1} \quad (14)$$

$$R_{ab} = (2\alpha) \left(\frac{n\alpha + N_a}{X_a} + \frac{n\alpha + N_b}{X_b} \right)^{-1} \quad (15)$$

The update formula is now

$$X_{ab} = Q_{ab} + R_{ab} \quad (16)$$

The key is that Q_{ab} is sparse, and R_{ab} has a simple functional form that is the result of vector operations. Furthermore, the iterative update does not require each R_{ab} , but rather $\sum_i R_{ib}$.

In practice, we find that this protocol remains limited by computational performance. As an alternative, the following regularization scheme appears to work well in practice.

Starting with the matrix C_{ij} of counts, we construct a matrix S_{ij} such that $S_{ij} = 1$ if $C_{ij} > 0$ or $C_{ji} > 0$. Thus, S is a sparse matrix with ones for every count that was observed in either the forward or reverse direction. When performing the MLE estimation, we use the matrix $C' = C + \alpha S$. The effect of this is to prevent transitions with limited statistics from being too strongly favored in one direction. In practice, α must be chosen such that $\alpha \sum_{ij} S_{ij} \leq \sum_{ij} C_{ij}$ for the data sets in this work, $\alpha \approx 0.1$ leads to $\alpha(\sum_{ij} S_{ij}) / (\sum_{ij} C_{ij}) \approx 0.01$. The advantages of this regularization are threefold. First, the data remain sparse, which allows scaling up to hundreds of thousands of states. Second, transitions that are nearly irreversible but inadequately sampled are smoothed. Third, this method adds pseudocounts only to transitions that were observed in the data (albeit in either the forward or reverse directions); thus, this method cannot introduce artificial pathways.

4. Two State Model for Comparing Transition Matrix Estimators. The two state model in Figure 4 is based on the transition matrix

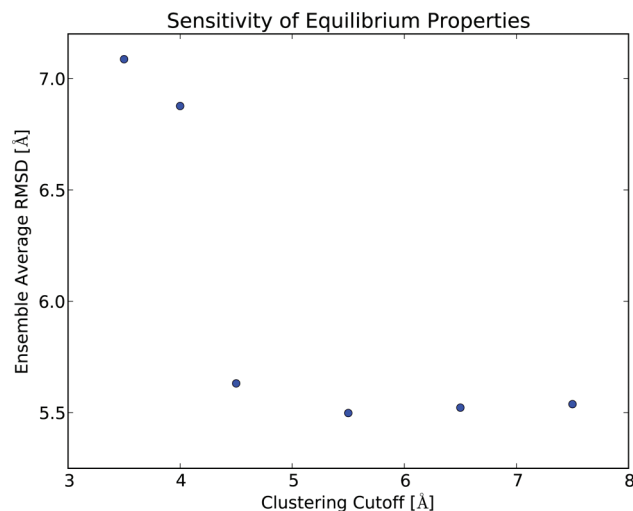


Figure 5. Ensemble average RMSD to native calculated for a sequence of models constructed from WW simulations.

$$T = \begin{pmatrix} p & 1 - q \\ 1 - p & q \end{pmatrix} \quad (17)$$

where $p = 0.99$ and $q = 0.999$. Thus, folding (100 timesteps) is approximately $10\times$ faster than unfolding (1000 timesteps); this is similar to the fast-folding variants of HP35⁴² under mildly denaturing conditions (with one timestep corresponding to 10 ns). Using this transition matrix, 100 trajectories of length 200 were generated and used to estimate transition and count matrices using either the symmetrized or reversible MLE protocols.

5. Balancing Kinetic Accuracy and Statistical Reliability.

Discretization error in MSM construction is reduced by increasing either the number of states or the lagtime.⁹ However, these solutions lead to statistical uncertainty due to increasing the number of model parameters or decreasing the amount of independent data, respectively. Thus, accurate model construction requires a careful balance between discretization and statistical error. A useful test is to consider the equilibrium properties of a sequence of models (Figure 5). We have calculated the ensemble average RMSD to native, which gives a smooth estimate of the stability of the folded state. For the WW protein, well-folded conformations typically show RMSD values of 0–4 Å, with unfolded conformations ranging from 5 to 10 Å. Models with few states ($f_{\max} \geq 4.5$ Å) appear near the folding midpoint, with an ensemble average RMSD of 5.54 ± 0.05 Å; models with more states ($f_{\max} = 3.5, 4.0$) appear considerably less folded, with a RMSD of 6.98 ± 0.1 Å. In general, state decompositions that are too fine will lead to spurious irreversible transitions and inaccurate equilibrium estimates. For the present data set (200 000 conformations), the 3.5 Å model has 47 684 states and lies well-within the data-poor regime. The lack of agreement with coarser models leads us to reject the 3.5 and 4.0 Å models. The 4.5 Å model is the best model for the WW data, as measured by relaxation time scale consistency (Figure 1a), correlation function analysis (Figure 1b), and equilibrium robustness (Figure 5). Constructing a sequence of models with increasingly many states helps identify models that minimize both discretization and statistical error.

6. Relaxation Time Scale Analysis of Alanine Dipeptide.

We present a relaxation time scale analysis (Figure 6) of a single

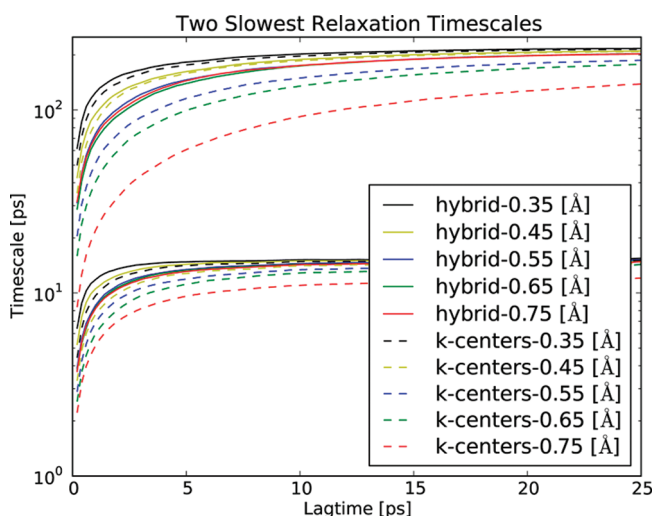


Figure 6. The two slowest relaxation time scales for alanine dipeptide are plotted as a function of lagtime.

(50 ns) alanine dipeptide simulation at 300 K in GBSA implicit solvent. In this example, the hybrid clustering provides improved performance for all choices of clustering diameter. Furthermore, the high-resolution models ($\epsilon \leq 0.45$ Å) converge to a slowest relaxation of 200 ps. The hybrid clusterings approach this value at shorter lagtimes, particularly for the lower-resolution models ($\epsilon \approx 0.65$ Å). The second-slowest time scale also suggests improved performance by the hybrid clustering.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Present Addresses

[†]Department of Molecular and Cell Biology, University of California, Berkeley, CA.

ACKNOWLEDGMENT

We thank Rhiju Das, Sergio Bacallado, and the members of the Das and Pande laboratories for helpful discussions. We gratefully acknowledge D. E. Shaw Research for providing the WW domain simulations. We thank NSF CNS-0619926 for computing resources and NIH R01-GM062868, NSF-DMS-0900700, and NSF-MCB-0954714 for funding. Finally, we thank Folding@Home donors for providing the computational resources for the HP35 simulation. K.A.B. was supported by a Stanford Graduate Fellowship. We acknowledge the following award for providing computing resources that contributed to the research results reported within this paper: MRI-R2: Acquisition of a Hybrid CPU/GPU and Visualization Cluster for Multidisciplinary Studies in Transport Physics with Uncertainty Quantification <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0960306>. This award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5). G. R. Bowman thanks the Miller Institute for funding.

REFERENCES

- (1) Inoue, A.; Saito, J.; Ikebe, R.; Ikebe, M. *Nat. Cell Biol.* **2002**, *4*, 302–306.
- (2) Anfinsen, C. *Science* **1973**, *181*, 223–230.
- (3) Buch, I.; Giorgino, T.; De Fabritiis, G. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184–10189.

- (4) Wüthrich, K. *J. Biol. Chem.* **1990**, *265*, 22059–22062.
- (5) Kendrew, J.; Bodo, G.; Dintzis, H.; Parrish, R.; Wyckoff, H.; Phillips, D. *Nature* **1958**, *181*, 662–666.
- (6) Kim, P.; Baldwin, R. *Annu. Rev. Biochem.* **1982**, *51*, 459–489.
- (7) Bai, Y.; Sosnick, T.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192–197.
- (8) Schuler, B.; Eaton, W. *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26.
- (9) Prinz, J.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105–174128.
- (10) Chodera, J.; Singhal, N.; Pande, V.; Dill, K.; Swope, W. *J. Chem. Phys.* **2007**, *126*, 155101–155118.
- (11) Bowman, G.; Beauchamp, K.; Boxer, G.; Pande, V. *J. Chem. Phys.* **2009**, *131*, 124101–124112.
- (12) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.
- (13) Voelz, V.; Bowman, G.; Beauchamp, K.; Pande, V. *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.
- (14) Lei, H.; Wu, C.; Liu, H.; Duan, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 4925–4930.
- (15) Buchete, N.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.
- (16) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.
- (17) Pan, A.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 064107–064115.
- (18) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105–204120.
- (19) Buchner, G. S.; Murphy, R. D.; Buchete, N.-V.; Kubelka, J. *Biochim. Biophys. Acta* **2011**, *1814*, 1001–1020.
- (20) Sarich, M.; Noé, F.; Schütte, C. *Multiscale Model. Simul.* **2010**, *8*, 1154–1177.
- (21) Scalco, R.; Cafilisch, A. *J. Phys. Chem. B* **2011**, *115*, 6358–6365.
- (22) Gonzalez, T. *Theor. Comput. Sci.* **1985**, *38*, 293–306.
- (23) Theobald, D. L. *Acta Crystallogr., Sect. A* **2005**, *61*, 478–480.
- (24) Kaufman, L.; Rousseeuw, P.; Corporation, E. *Finding groups in data: an introduction to cluster analysis*; Wiley Online Library: New York, 1990; Vol. 39.
- (25) Keller, B.; Daura, X.; van Gunsteren, W. *J. Chem. Phys.* **2010**, *132*, 074110–074126.
- (26) Tarjan, R. *SIAM J. Comput.* **1972**, *1*, 146–160.
- (27) Bacallado, S.; Chodera, J.; Pande, V. *J. Chem. Phys.* **2009**, *131*, 045106–045116.
- (28) Diaconis, P.; Rolles, S. *Ann. Stat.* **2006**, *34*, 1270–1292.
- (29) Cellmer, T.; Buscaglia, M.; Henry, E.; Hofrichter, J.; Eaton, W. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6103–6108.
- (30) Haque, I.; Beauchamp, K.; Pande, V. Submitted 2011.
- (31) Maiorov, V. N.; Crippen, G. M. *J. Mol. Biol.* **1994**, *235*, 625–634.
- (32) Damm, K.; Carlson, H. *Biophys. J.* **2006**, *90*, 4558–4573.
- (33) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011–19016.
- (34) Rossum, G. *Python reference manual*; CWI (Centre for Mathematics and Computer Science): Amsterdam, The Netherlands, 1995.
- (35) Ascher, D.; Dubois, P. F.; Hinsen, K.; Hugunin, J.; Oliphant, T. *Numerical Python*, version UCRL-MA-128569; Lawrence Livermore National Laboratory: Livermore, CA, 1999.
- (36) Alted, F.; Vilata, I. <http://www.pytables.org/> (accessed June 1, 2011).
- (37) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- (38) Jager, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M.; Noel, J.; Gruebele, M.; Kelly, J. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 10648–10653.
- (39) Peng, T.; Zintsmaster, J.; Namanja, A.; Peng, J. *Nat. Struct. Mol. Biol.* **2007**, *14*, 325–331.
- (40) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.; Dror, R.; Shaw, D. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.
- (41) Bowman, G.; Ensign, D.; Pande, V. *J. Chem. Theory Comput.* **2010**, *6*, 787–794.
- (42) Kubelka, J.; Chiu, T.; Davies, D.; Eaton, W.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546–553.

Hybrid Quantum and Classical Simulations of the Dihydrofolate Reductase Catalyzed Hydride Transfer Reaction on an Accurate Semi-Empirical Potential Energy Surface

Dvir Doron,[†] Dan Thomas Major,^{*,†} Amnon Kohen,[‡] Walter Thiel,[§] and Xin Wu[§]

[†]Department of Chemistry, The Lise Meitner-Minerva Center of Computational Quantum Chemistry, Bar-Ilan University, Ramat-Gan 52900, Israel

[‡]Department of Chemistry, University of Iowa, Iowa City, Iowa 52242, United States

[§]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany

S Supporting Information

ABSTRACT: Dihydrofolate reductase (DHFR) catalyzes the reduction of 7,8-dihydrofolate by nicotinamide adenine dinucleotide phosphate hydride (NADPH) to form 5,6,7,8-tetrahydrofolate and oxidized nicotinamide. DHFR is a small, flexible, monomeric protein with no metals or SS bonds and serves as one of the enzymes commonly used to examine basic aspects in enzymology. In the current work, we present extensive benchmark calculations for several model reactions in the gas phase that are relevant to the DHFR catalyzed hydride transfer. To this end, we employ G4MP2 and CBS-QB3 ab initio calculations as well as numerous density functional theory methods. Using these results, we develop two specific reaction parameter (SRP) Hamiltonians based on the semiempirical AM1 method. The first generation SRP Hamiltonian does not account for dispersion, while the second generation SRP accounts for dispersion implicitly via the AM1 core-repulsion functions. These SRP semiempirical Hamiltonians are subsequently used in hybrid quantum mechanics/molecular mechanics simulations of the DHFR catalyzed reaction. Finally, kinetic isotope effects are computed using a mass-perturbation-based path-integral approach.

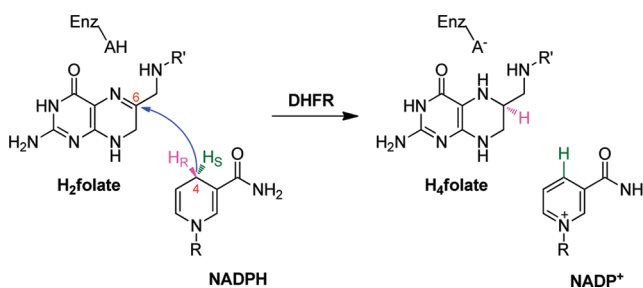
INTRODUCTION

Dihydrofolate reductase (DHFR; EC 1.5.1.3) catalyzes the reduction of 7,8-dihydrofolate (H_2 folate) by nicotinamide adenine dinucleotide phosphate hydride (NADPH) to form 5,6,7,8-tetrahydrofolate (H_4 folate) and $NADP^+$. Its principal function is to maintain intracellular pools of H_4 folate, which in turn serves as a cofactor in one-carbon metabolic processes and is essential for the biosynthesis of purines, thymine nucleotides, and several amino acids. DHFR has long been recognized as an important target for various therapeutic purposes, in particular the development of anticancer and antibacterial drugs, such as methotrexate and trimethoprim, respectively.^{1,2} The clinical importance of DHFR, along with its relatively modest size (159 aa in *E. coli* DHFR), has led many researchers to study, both experimentally and theoretically, the catalytic mechanism and kinetics of the NADPH-dependent hydride transfer reaction.¹

The key chemical step in the catalytic cycle of DHFR involves a stereospecific transfer of the *pro-R* hydrogen at the C4 position of the nicotinamide ring in NADPH to the *si*-face of the C6 atom of the pterin ring in H_2 folate, with concomitant protonation at the adjacent N5 position (Scheme 1).^{3,4}

Early kinetic studies of *E. coli* DHFR by Fierke et al. established a catalytic pathway cycling between five intermediates, including the $E \cdot NADPH$, the Michaelis complex $E \cdot NADPH \cdot H_2$ folate, the ternary product complex $E \cdot NADP^+ \cdot H_4$ folate, the binary product complex $E \cdot H_4$ folate, and the product release complex $E \cdot NADPH \cdot H_4$ folate.⁵ At neutral pH, the hydride-transfer rate in the wild-type enzyme is 220 s^{-1} and the rate-determining step is the release of the product H_4 folate (12 s^{-1}),

Scheme 1. Hydride Transfer Reaction Catalyzed by DHFR^a



^a R: adenine dinucleotide 2'-phosphate. R': *p*-aminobenzoyl-glutamate.

whereas at high pH, the hydride transfer step becomes more rate limiting, suggesting that the protonated substrate (henceforth H_3 folate⁺) is the reactive species for the hydride transfer reaction.⁵ Thus, the traditional view is that the protonation step precedes hydride transfer, thereby generating a resonance-stabilized iminium intermediate that serves as a more efficient acceptor toward the negatively charged hydride ion.^{4,6–8} Indeed, vibrational spectroscopic studies by Callender et al. have demonstrated that the N5 pK_a of H_2 folate is raised from 2.6 in solution to 6.5 in the ternary product complex with *E. coli* DHFR.^{9–12} The elevation of the N5 pK_a by four units upon complex formation is likely an enzymatic strategy for a substantial rate enhancement

Received: July 10, 2011

Published: August 26, 2011

over the uncatalyzed reaction in solution. It is likely that the hydrophobic nature of the active site pocket maintains a relatively low permittivity, thereby lowering $\text{H}_3\text{folate}^+$ acidity. Nevertheless, the protonation source is a rather controversial issue: originally, it was believed to be Asp27, being the only ionizable residue within the binding site of *E. coli* DHFR.^{10,13,14} However, the studies of Callender and co-workers suggested that the N5 atom is responsible for the pH dependency of the reaction and that the Asp27 residue exists in a deprotonated form at physiological pH and does not donate a proton to the substrate during enzymatic catalysis. It has been suggested instead that the negative charge of the carboxylate of Asp27 could stabilize the protonated substrate even though this group is on the other side of the bound substrate relative to N5.^{10,11,15} This notion was supported by computational studies of Brooks and Rod, arguing that the protonation of the substrate's N5 position comes directly from the bulk solvent.¹⁶

From a structural point of view, DHFR is a small monomeric protein (~ 18 kD, 159 aa for the *E. coli* DHFR).^{17–20} The substrate and coenzyme bind in a deep hydrophobic cleft at the juncture of the adenosine binding subdomain and the major (“loop”) subdomain. In its “closed” conformation, the Met20 loop (residues 9–24) lies directly over the active site, shielding the reactants from the solvent, and is primarily responsible for determining the active site architecture. X-ray studies in numerous ligand-bound states show that the Met20 loop assumes four characteristic conformations with respect to the active site.²⁰ In particular, the movement of Met20 loop and the cycling of the $\beta\text{F}-\beta\text{G}$ and $\beta\text{G}-\beta\text{H}$ loops between the *closed* and *occluded* conformations are coordinated with the stages of the catalytic cycle.²⁰

Early computational studies on the hydride transfer reaction catalyzed by DHFR were typically performed for small model systems *in vacuo*.^{21–24} In these studies, the potential energy surface (PES) was traditionally explored at the semiempirical or *ab initio* levels of theory. Subsequently, hybrid quantum-mechanical/molecular-mechanical (QM/MM) studies of DHFR have enabled the examination of environmental effects.^{4,25–27} Indeed, Moliner and co-workers in their pioneering QM/MM study on DHFR⁴ illustrated the importance of incorporating the enzyme's degrees of freedom in the study of transition state structures, compared to those calculated for the gas phase models.⁴ However, the calculated free energies of reaction and activation barriers by this approach were unrealistic compared to the experimental data, overestimating the free energy barrier for the enzymatic reaction, due to the inaccurate representation of the atoms in the QM region by a semiempirical method.

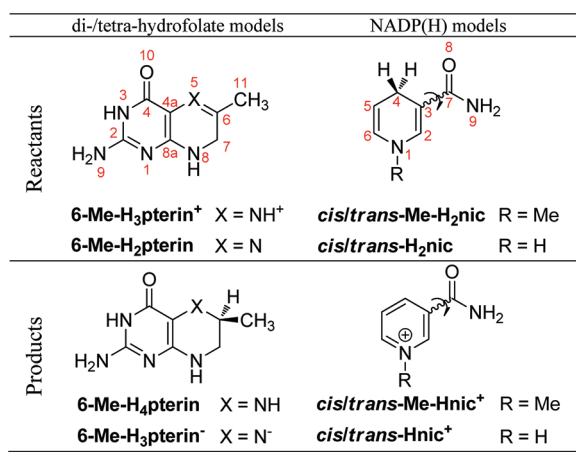
Nevertheless, numerous research groups took advantage of the computational efficiency of the hybrid semiempirical QM/MM approach combined with molecular dynamics (MD) simulations, to explore the hydride transfer reaction in DHFR^{7,28–30} and other related dehydrogenases.^{31–34} Some of these works^{7,28,30} attempted to compensate for the inherent errors in the semiempirical models by introducing correction terms. In particular, Gready and co-workers derived correction terms for selected configurations estimated from analogue cluster calculations at the DFT and post-Hartree–Fock theoretical levels, and these were applied to the activation free energy.²⁸ However, the authors did not manage to obtain quantitative agreement with the experimentally observed free energy barrier. On the other hand, Garcia-Viloca et al.,⁷ and more recently Brooks and Thorpe,³⁰ enhanced the semiempirical Hamiltonian with a simple valence bond (SVB) correction term,³⁵ which was parametrized to fit two quantities: the activation free energy of 13.4 kcal/mol,

which is predicted by transition-state theory expression taking the experimental pH-independent hydride transfer rate constant (950 s^{-1} at $25 \text{ }^\circ\text{C}$)⁵ into account, and the reaction free energy of -4.4 kcal/mol, calculated from the reported equilibrium constant (1700).⁵ The disadvantage of such an approach is that for systems where the original QM method performs poorly, the SVB term is necessarily large. This may introduce artifacts into the vibrational frequency related to the SVB (reaction) coordinate, as well as an artificial increase in energy at the end points of the reaction coordinate.⁷ Such QM/MM potentials have been employed in combination with ensemble averaged variational transition state theory with multidimensional tunneling (EA-VTST/MT) calculations to obtain kinetic isotope effects.³⁶

A different QM/MM approach to studying hydride transfer reactions is the empirical valence bond (EVB) potential for the QM region, which was developed by Warshel and Weiss^{37,38} and parametrized for DHFR by Hammes-Schiffer et al.^{39,40} and Warshel and Liu.⁴¹ In this method, the hydride transfer reaction is represented by means of two empirical valence bond states, namely, reactants and products. The matrix elements between these diabatic states are represented as MM terms which are parametrized to reproduce experimental free energies of reaction and activation (dictated by either the pH-independent³⁹ or -dependent⁴⁰ intrinsic rate constants). An advantage of the EVB approach is that the reaction coordinate can naturally be expressed in terms of a collective entity (i.e., the energy gap between the reactant and product diabatic states). Additionally, the EVB method is computationally very efficient. On the other hand, due to its simplistic form and depending on the parametrization strategy, the fine details of the potential energy surface may not be described correctly. Furthermore, the experimental assessment of intrinsic rates in a complex kinetic cascade is limited, and the presumed experimental rate constant is commonly a complex kinetic term with many microscopic rate constants that cannot be distinguished experimentally. Therefore, using experimentally determined rate constants to parametrize the EVB terms might be problematic. The EVB potential has been employed to incorporate nuclear quantum effects by representing the transferring hydrogen nucleus as a 3-D vibrational wave function,³⁹ as well as in Feynman path-integral (PI) simulations.⁴² For a tabulated summary of some prominent computational studies on the DHFR hydride transfer reaction published in the past decade, the reader is referred to the Supporting Information.

To the best of our knowledge, no simulation of DHFR has gone beyond the EVB level or the semiempirical (AM1/PM3) Hamiltonian levels, using the standard parameters. In this paper, we first present gas-phase model calculations for the hydride transfer reaction between H_2folate and NADPH, using high-level *ab initio* and density functional theory methods. Although these models lack the contribution of the enzymatic environment, the calculations shed light on some key thermodynamic aspects related to the intrinsic thermochemistry of the reaction. Second, we present an accurate potential energy surface for the hydride transfer reaction in the enzyme *E. coli* DHFR, taking advantage of the comprehensive gas phase calculations presented herein. This potential energy surface is described by a hybrid quantum mechanics/molecular mechanics (QM/MM) potential, where the QM subsystem is treated by a semiempirical model that has been specifically parametrized to reproduce *ab initio* and DFT data. Consequently, the quality of such a specific reaction parameter (SRP) model^{43–45} is comparable with calculations at the *ab initio* and DFT levels, but at a considerably lower cost, hence

Chart 1. Main Models Subjected to Gas-Phase Calculations in This Study



allowing us to perform long MD simulations with the full solvated enzyme. The calibration of the AM1 Hamiltonian was performed only for the reaction in the gas phase (i.e., the intrinsic performance of the QM model), while the interactions with the environment are captured via the QM/MM interaction terms. Subsequently, the performance of the QM/MM potential was validated for the enzymatic reaction. This is in contrast to models employed in earlier studies, which parametrize empirical QM models such as EVB and SVB to the experimental free energies of reaction and activation in solution or in the enzyme. Two principle SRP models were explored, differing in the way they treat dispersion interactions. Nuclear quantum effects (NQE) are described using a coupled free-energy mass-perturbation and umbrella sampling simulation technique employing Feynman centroid path integral calculations (PI-FEP/UM).⁴⁶ Thus, both the electronic structure of the reacting system and the nuclear dynamics are treated quantum mechanically. This method has been demonstrated in a series of studies of chemical reactions in solution and in enzymes.^{44–48}

METHODOLOGY

Gas Phase QM Calculations. Model Reactions. The molecules depicted in Chart 1 represent chemical analogues of the reacting ligands and their corresponding products involved in the DHFR enzymatic reaction. The geometries were fully optimized in the gas phase using the Gaussian 09 program⁴⁹ and seven different density functionals, including one generalized gradient approximation (GGA) functional [PBE/PBE (PBE)⁵⁰], two hybrid GGA functionals [PBE1PBE (PBE0)⁵¹ and B3LYP^{52–54}], and four hybrid meta-GGA functionals (B98,⁵⁵ BBIK,⁵⁶ MPWB1K,⁵⁶ and M06⁵⁷). For any functional, the 6-31+G(d,p) basis set was found to give adequately converged geometries and reaction energies (results are not shown for other basis sets). The same geometries were recalculated with the semiempirical potential Austin Model 1 (AM1),⁵⁸ using the standard parameter scheme within Gaussian 09. We also carried out high-level calculations with the complete basis set method CBS-QB3,^{59,60} as well as the Gaussian theory method G4MP2.^{61,62} The latter method provided most of the target values for reparametrization of the AM1 Hamiltonian (*vide infra*).

The H₂folate substrate and its corresponding H₄folate product are represented by derivatives of 6-methyl-7,8-dihydropterin and

6-methyl-5,6,7,8-tetrahydropterin. These molecular models are the substrate and product of the DHFR catalyzed reaction, and it is assumed that the mechanism of reduction of the 7,8-dihydropterin is similar to that of 7,8-dihydrofolate.⁶³ To investigate the effect of protonation of the pyrazine's N5 atom on thermodynamics, two model reactions were inspected: the “protonated” model, which involves the protonated form of the dihydropterin reactant (6-Me-H₃pterin⁺), ending up in a neutral reduced tetrahydropterin product (6-Me-H₄pterin), and the “unprotonated” model, in which the unprotonated dihydropterin (6-Me-H₂pterin) is reduced to a negatively charged species (6-Me-H₃pterin⁻).

Two conformers of the nicotinamide derivative representing the NADP(H) cofactor were considered, distinguished by the orientation of the carboxamide with respect to the (dihydro)pyridine ring: the *cisoid* conformer, in which the carbonyl and the C2=C3 bond are quasi-synperiplanar, and the *transoid* conformer, where these two are quasi-antiperiplanar. As the two types of conformers were taken into account for both the reduced and oxidized nicotinamide species (Me-H₂nic and Me-Hnic⁺, respectively), four thermodynamic pathways were computed for each model reaction separately, as depicted in Scheme 2.

In both model reactions, the changes in electronic energy (ΔE_{el}), enthalpy (ΔH), and Gibbs free energy (ΔG) at 298 K were computed for each pathway with the aforementioned methods, according to the following general equations:

$$\Delta X_r^{\text{prot}} = X[6\text{-Me-H}_4\text{pterin}] + X[\text{cis/trans-Me-Hnic}^+] - X[6\text{-Me-H}_3\text{pterin}^+] - X[\text{cis/trans-Me-H}_2\text{nic}] \quad (1a)$$

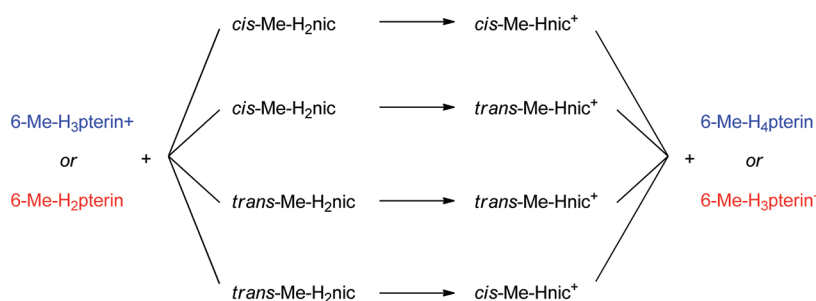
$$\Delta X_r^{\text{unprot}} = X[6\text{-Me-H}_3\text{pterin}^-] + X[\text{cis/trans-Me-Hnic}^+] - X[6\text{-Me-H}_2\text{pterin}] - X[\text{cis/trans-Me-H}_2\text{nic}] \quad (1b)$$

where X is a general notation for E_{el} , H , or G , and the superscripts X^{prot} and X^{unprot} refer to the protonated and unprotonated model reactions, respectively.

Modeling the Reactant, Transition, and Product State Complexes. The optimized model structure for the transition state complex was found using the synchronous transit-guided quasi-Newton (STQN) method⁶⁴ implemented in the Gaussian 09 program,⁴⁹ with all seven density functionals and the standard semiempirical AM1 potential (the G4MP2 method turned out to be too costly for transition state optimization of the bimolecular complex). In the case of the B3LYP functional, an empirical Grimme-type dispersion correction was also added (denoted B3LYP-D).^{65,66} The *transoid* conformer of the Me-H₂nic subunit was chosen, because it was found to be the most prevalent conformer identified in X-ray crystal structures of most enzyme active sites,^{67,68} in particular *E. coli* DHFR.²⁰ The saddle point was identified by a single imaginary vibrational frequency corresponding to the normal mode of transferring the hydride between the donor (C4 in the nicotinamide subunit) and acceptor (C6 in the pterin subunit) carbons.

The reactant and product complexes, [6-Me-H₃pterin⁺ · *trans*-Me-H₂nic] and [6-Me-H₄folate · *trans*-Me-Hnic⁺], respectively, were obtained by intrinsic reaction coordinate (IRC) calculations in the direction of reactants and products. The steepest descent path in mass-scaled coordinates was followed using 100 steps of 0.1 Bohr in each direction of the reaction path

Scheme 2. Possible Thermodynamic Pathways for the Protonated (Blue) and Unprotonated (Red) Model Reactions of the Hydride Transfer in the Gas Phase



down to the reactant and product complex wells, where the root-mean-square gradient norms at the end points were no higher than $0.2 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$ and $0.3 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$ in the forward and reverse directions, respectively. Wherever methods accounting for dispersion (B3LYP-D and M06) were used, the IRC calculations were performed further until the default convergence criteria were reached in each IRC direction. The IRC end points were further geometry optimized as was done for the individual molecules.

Development of Semiempirical Specific Reaction Parameters. In the SRP approach,⁴³ the semiempirical parameters are optimized for a given system, i.e., here to treat the hydride transfer reaction in DHFR. More specifically, AM1-SRP parameters were developed to reproduce electronic and thermodynamic properties obtained from high-level QM calculations on representative molecular models in the gas phase (*vide supra*). For the majority of the modeled species, the reference method for the AM1 parametrization was the composite Gaussian method G4MP2,^{61,62} whereas target quantities associated with models of the reactant, transition, and product state bimolecular complexes were calculated using the B3LYP, B3LYP-D, and M06 hybrid functionals^{52–54} with the 6-31+G(d,p) basis set. The observables used as target values were enthalpies of formation, reaction energies, geometries, dipole moments, Mulliken charges, and vibrational frequencies.

Briefly, the current SRP strategy aims at two major goals: (1) an electronic PES which is of high accuracy, comparable to a high-level *ab initio* or DFT PES, and (2) absolute atomization energies which give rise to accurate heats of formation. The former goal allows the SRP model to be employed in MD simulations where all classical thermal effects are included directly via the propagation of Newton's equations of motion, while all quantum thermal effects are included via path-integral simulations or similar methods. The latter objective serves to limit deviations from the original AM1 parameter set (which has been optimized to reproduce heats of formation) to a minimum, as many basic molecular properties are treated well by AM1, although energy values are often not sufficiently accurate for quantitative comparison with experimental results. The SRP parameters are obtained by a nonlinear optimization, starting with the original AM1 parameters as the initial input.⁶⁹ The following general sequential optimization scheme is adopted: (1) First, optimize the one-center energies (U_{ss} and U_{pp}) and the resonance integrals (β_s and β_p), followed by the α parameters in the core-repulsion function (CRF). (2) Adjust the orbital exponents (ζ_s and ζ_p) together with the previously optimized parameters. (3) Adjust the one-center two-electron repulsion integrals G_{ss} , G_{sp} , G_{pp} , G_{p2} (or $G_{pp'}$),

and H_{sp} together with the previously optimized parameters. (4) Include the Gaussian CRF parameters L , M , and K together with the previously optimized parameters.

The training set for the optimizations consisted of the 12 individual molecules in Chart 1, plus the neutral nicotinamides, *trans*-nic and *cis*-nic. Additionally, the reactant, product, and transition states and selected structures along the IRC reaction paths were also included. All individual molecules were fully geometry optimized during the SRP parametrization, while the complexes were either partially or fully optimized. Single point calculations were carried out for the structures along the IRC.

In order to assess the importance of dispersion interactions, we attempted to implicitly include dispersion effects into the AM1-SRP model via the AM1 CRFs. In this approach, an improved SRP Hamiltonian was developed which implicitly accounts for dispersion. The motivation for including dispersion was provided by initial attempts to optimize the geometry of the reactant and product complexes using standard AM1. This led to unrealistic gas-phase structures, presumably due to the lack of dispersion interactions in standard AM1; for example, the two fragments in the product complex moved far apart (see Results and Discussion section). In order to capture the dispersion interactions implicitly, an empirical Grimme-type dispersion correction was added to the CRF in standard AM1, and thereafter all parameters in the Gaussian AM1 terms were reoptimized to best fit this dispersion-corrected CRF by using a nonlinear least-squares procedure (see Appendix A for technical details). The original CRF, the added dispersion corrections, the sum of these two terms, and the best fit Gaussian terms for the H–H, C–C, N–N, and O–O atom pairs are plotted in Figure 1. The sum and fitted curves for the first three pairs are visually almost indistinguishable, and there are merely minor variations of the standard AM1 CRF parameters (Figure 1). In the case of the O–O pair, the fitted Gaussian terms for oxygen, however, are not able to reproduce the attractive well in the van der Waals region due to the presence of two positive (repulsion) Gaussian functions ($K > 0$). Since oxygen is chemically less important for the DHFR reaction, the standard AM1 parameters for oxygen in the Gaussian terms were retained except in the very last step (4) of the SRP parametrization. The dispersion corrected CRF parameters for hydrogen, carbon, and nitrogen were used in the subsequent optimization procedure described above (fixed in steps 1–3, and adjusted in step 4).

Following the strategy outlined above, two SRP Hamiltonians were developed and employed in the simulations on DHFR. Specifically, an initial model termed AM1-SRP was designed, where the target complex structures and properties employed in the SRP development process were obtained from B3LYP.

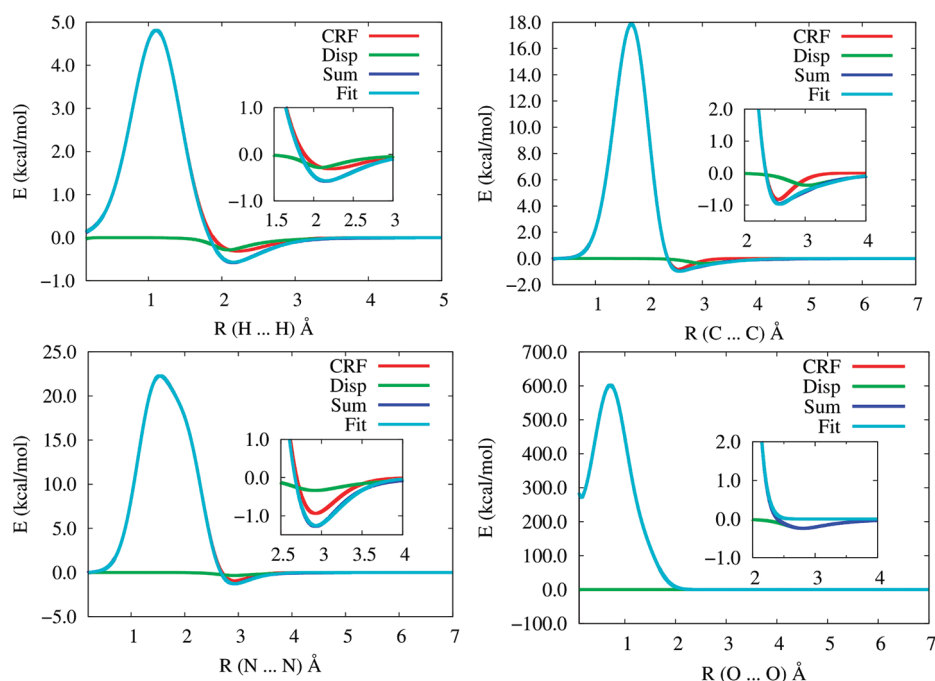


Figure 1. The nonlinear least-squares curve fittings for pairs of hydrogen, carbon, nitrogen, and oxygen atoms. CRF is the standard AM1 core repulsion function. Disp is the dispersion energy correction calculated using Grimme's formula. Sum is the total energy of CRF and Disp. Fit is the Gaussian term fitted with respect to Sum. A magnified subplot of the van der Waals region is also shown for each atom pair.

In this AM1-SRP model, intermolecular dispersion effects are not included during the parametrization process. Subsequently, a SRP Hamiltonian which implicitly includes dispersion was devised [termed AM1-SRP(D)]. Here, all complex structures were obtained from B3LYP-D calculations. In both approaches, the target data for the individual molecules were taken from G4MP2 calculations (accounting for intramolecular dispersion) at B3LYP optimized geometries.

Target values are presented in the Supporting Information, along with a detailed explanation of their determination, validation of the AM1-SRP performance, and a complete list of the modified parameters. The AM1-SRP results presented in the Supporting Information were obtained using the Gaussian 09 program.

Modeling QM/MM Interactions. To investigate the ability of the developed AM1-SRP Hamiltonians to accurately model QM/MM interactions, we computed the complexation energies between selected QM molecules and a TIP3P water molecule. The QM molecules are 6-Me-H₃pterin⁺, *trans*-Me-H₂nic, 6-Me-H₄pterin, and *trans*-Me-H₂nic⁺. For each of these QM moieties, a single TIP3P water molecule was placed at different hydrogen bonding positions around the molecule, for a total of up to four complexes per QM molecule. The QM/MM complexation energies were compared to complexation energies computed using various high-level methods.

QM/MM Simulations of *E. coli* DHFR. *Model of the Ternary Complex of *E. coli* DHFR.* The crystal structure of *E. coli* DHFR with folate and the oxidized cofactor NADP⁺ (PDB ID code: 1rx2), originally reported by Sawaya and Kraut,²⁰ was used to construct the initial configuration for the present study. This structure corresponds to the Michaelis complex of *E. coli* DHFR with the Met20 loop in the closed conformation. The X-ray crystal structure contains a total of 159 amino acid residues, 153 crystallographic waters, and the folate and NADP⁺ ligands,²⁰ which were replaced by H₂folate and NADPH, respectively, for

the simulation. The N5 atom on the substrate's pyrazine ring was protonated, as this form is thought to be the active species for the hydride transfer reaction.^{11,70}

On the basis of the results of Callender and co-workers,^{10,11,15} Asp27 was assumed to be deprotonated at neutral pH, while the protonation states for all other ionizable residues were set corresponding to pH 7.^{5,71,72} The hydrogen bonding patterns of the ionizable residues with the surrounding environment were visually inspected to verify that the protonation states are reasonable. The coordinates of hydrogen atoms of the protein, water, and coenzyme were determined using the HBUILD facility in the program CHARMM.^{73,74} The possible protonation states of histidine residues (proton on Ne, proton on Nd, or doubly protonated form) were determined by examination of the hydrogen bonding interactions. Peripheral/surface His residues were generally assumed to be positively charged. In NADPH, the 2'-phosphate moiety on the adenosine ribose was treated as a dianion, based on a pK_a of ~5.9 reported for NADP⁺ bound in cytochrome P-450 oxidoreductase (P-450R).⁷⁵ The coordinates of NE2 and OE1 atoms of the carboxamide moiety in the Gln102 residue were swapped, with respect to the original solved crystal structure, in order to form plausible hydrogen bond interactions with the adenine group of NADPH (a similar analysis was carried out by Brooks et al.⁷⁶). Residue 37 was built as Asp37 to be consistent with the 1RX2 PDB entry and previous work on DHFR, although sequencing data suggest that this residue might be Asn37.^{77,78} We note that the DHFR function is insensitive to the nature of the amino acid at position 37.⁷⁹ The resulting negatively charged enzyme (−14) has dimensions of ca. 34 × 42 × 50 Å³. To this system, we added 14 sodium ions in random positions outside the protein to obtain a net-neutral system, a prerequisite for convergence of the Ewald summation method (*vide infra*).⁸⁰ Subsequently, the protein, ligands, crystal waters, and counterions were embedded in a water box as detailed below.

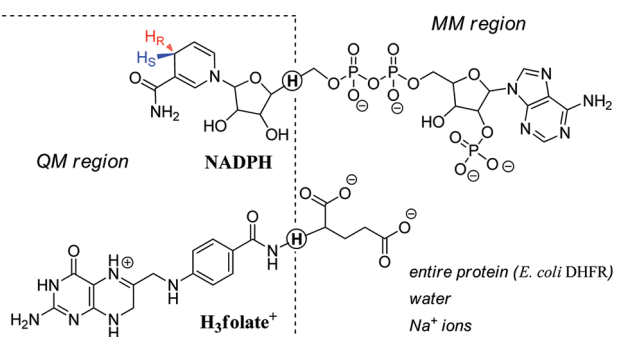


Figure 2. QM/MM partitioning scheme. The dashed line divides the QM and MM regions, and the quantum hydrogen link atoms are circled.

Hybrid QM(SRP)/MM Potential Energy Surface. The hydride transfer reaction in *E. coli* DHFR was described using a hybrid QM/MM potential energy surface.

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \quad (2)$$

The system was partitioned into a QM region consisting of 69 atoms and a MM region containing the rest of the system. The QM subsystem includes 38 atoms from the $\text{H}_3\text{folate}^+$ substrate (the pterin ring, the *N*-methylene-substituted *p*-aminobenzoyl (pABA) moiety, and the NH atoms of the glutamate moiety) and 29 atoms from the NADPH coenzyme (the dihydronicotinamide and ribose rings). In addition, two hydrogen link atoms were introduced along the covalent bonds crossing the boundary between the QM and the MM regions, to satisfy the valence requirements of the QM fragments. A schematic representation is depicted in Figure 2, where the quantum link atoms are circled.

The QM region was treated by the AM1-SRP or the AM1-SRP(D) Hamiltonians described above. The all-atom CHARMM22 force field⁸¹ with grid-based energy correction maps (CMAP)⁸² for peptide dihedral angles was employed to treat the entire protein, the substrate, and the ions, while the CHARMM27 force field⁸¹ was used for the coenzyme. The water molecules were represented by the three-point charge TIP3P model.⁸³ Atom types and parameters for the substrate were assigned by analogy with existing functional groups, as reported by Garcia-Viloca et al.⁷ For structure minimization and initial equilibration at the pure MM level, partial atomic charges for the substrate were assigned using the CHARMM force field (Momany-Rone) integrated in Discovery Studio 2.5 (Accelrys Software Inc., San Diego, CA).

QM/MM interactions were treated by electrostatic embedding wherein the MM partial atomic charges are included in the one-electron Hamiltonian. To fine-tune the PES, QM/MM interaction energies between the reacting fragments (QM) and the protein (MM) were modified. The van der Waals (vdW) parameters of the QM hydrogen atoms were changed to those that reproduce the interaction energies for hydrogen bonded complexes in the gas phase obtained from *ab initio* calculations at the MP2/6-31G(d,p) level.⁸⁴

*Free-Energy Simulations—General Approach.*⁴⁶ We follow a two-step procedure⁸⁵ in which we first carry out Newtonian MD simulations to determine the classical mechanical potential of mean force (PMF) along the reaction coordinate for the hydride transfer reaction between $\text{H}_3\text{folate}^+$ and NADPH in the fully solvated enzyme. Then, atoms that are directly involved in the hydride transfer are quantized, and the configurations sampled in

MD simulations are used in path-integral simulations by constraining the centroid positions of the quantized particles to the classical coordinates. This double (quantum and classical) averaging scheme is formally rigorous^{46,85–88} and yields the QM-PMF as a function of the centroid reaction path.^{89,90} In PI-FEP/UM, the ratio of the quantum partition functions for different isotopes, which yields the kinetic isotope effects (KIEs), is obtained by free-energy perturbation from a light isotope mass into a heavier one within the same centroid path-integral simulation,⁴⁶ avoiding the difference between two free-energy barriers with greater fluctuations than the difference itself for the two isotopic reactions. Consequently, the PI-FEP/UM method is unique in that it yields accurate results for computed KIEs, including secondary KIEs.^{45,46}

MD Simulations. MD simulations were conducted under periodic boundary conditions (PBC), with Ewald summation for electrostatic interactions.⁹¹ The solute was soaked in a pre-equilibrated $65 \times 65 \times 65 \text{ \AA}^3$ cubic box of 9461 water molecules, with its longest axis lying along the space diagonal of the box to ensure that all protein atoms are at least 10 \AA away from the edges of the box. The final model contained 27 986 atoms. For van der Waals and electrostatic interactions, a 13.0 \AA group-based cutoff was used. The Ewald method was employed for reciprocal space summations between MM sites as well as for the QM/MM interactions using a $64 \times 64 \times 64$ FFT grid.⁹¹ The κ value was set to 0.340 \AA^{-1} .

All water molecules were relaxed using the adopted-basis set Newton–Raphson (ABNR) minimization method (30 steps), while the crystal water oxygens were harmonically restrained to their original positions. This was followed by a 100 ps MD equilibration of the water molecules, which were thereafter minimized again (30 steps ABNR). Afterward, all atoms were subjected to minimization in a stepwise fashion, to remove close contacts in the initial protein–ligand–solvent system: (a) The substrate and coenzyme molecules were first minimized (30 step ABNR) while placing harmonic restraints on heavy atoms and keeping the rest of the system fixed. The restraints were gradually decreased to zero, while the ligands were further minimized (5×30 steps ABNR). (b) The water molecules and protein molecules were minimized (this time the ligands were held fixed) while the harmonic restraints on their heavy atoms were gradually diminished (4×10 steps ABNR). (c) Eventually, the whole system was minimized (30 steps ABNR) without any restraints.

The isothermal–isobaric ensemble (NPT) was employed at 1 atm and 298 K using the extended system pressure/temperature (CPT) algorithm of Andersen⁹² with an effective mass of 500 amu and a Hoover thermostat⁹³ with an effective mass of 1000 kcal/mol·ps². The SHAKE algorithm⁹⁴ was applied to constrain all MM bonds involving hydrogen atoms, allowing a time-step of 1 fs. The system was gradually heated up from 48 to 298 K during five sessions of 5 ps for a total of 25 ps and thereafter equilibrated at the target temperature (298 K) over the course of 1 ns at the MM level of theory, with a further 200 ps of equilibration using the QM(SRP)/MM potential.

In light of the flexibility of the protein and the structural manipulation of the original ligands bound in the crystal structure, some issues emerged during the equilibration phase which required intervention. This included transient introduction of nuclear Overhauser effect (NOE) harmonic restraints on the distance between hydride donor and acceptor carbons (C4N in NADPH and C6 in $\text{H}_3\text{folate}^+$, respectively), as well as on other distances between donors and acceptors of selected hydrogen bonds within the protein which are characteristic of the closed

conformation (for details, see the Supporting Information). All restraints were removed 100 ps prior to commencing the production phase.

Potential of Mean Force. The classical-mechanical potential of mean force (CM-PMF)⁹⁵ was determined using the umbrella sampling technique, in order to sample the high-energy regions of the potential energy surface.⁹⁶ The reaction coordinate (ζ) was defined geometrically as the difference between the lengths of the breaking (C4N_{NADPH}-H4N) and forming (H4N-C6_{H₃folate}) bonds. A total of 13 discrete regions along the reaction coordinate (“windows”) were defined with a uniform spacing of 0.25 Å. Each simulation was performed with the addition of a biasing potential (roughly the negative of the computed PMF), and a harmonic restraint centered at each window. The harmonic force constants, k , ranged between 20.0 and 60.0 kcal·mol⁻¹·Å⁻² [$E_{\text{harm}} = k(\zeta - \zeta_{\text{ref}})^2$]. Each window was equilibrated for 2 ps, followed by a 100 ps production simulation that collected the probability densities of configurations (ρ) along the reaction coordinate (ζ) and sorted them into bins of width 0.01 Å. The coordinates were saved every 0.5 ps, and the velocities and positions of the last configuration generated in each window were used to initiate the next window. The PMF curve was obtained using the weighted histogram analysis method (WHAM).⁹⁷ To ensure convergence of the PMF, the simulations were run until the difference between sequential PMF profiles was less than ± 1 kcal/mol. The QM-PMF was obtained using a double averaging procedure by centroid path-integral simulations on configurations saved during the umbrella sampling.^{46,85,87} In essence, the centroid path-integral simulations yield the free energy difference between the classical mechanical and the quantum mechanical PMFs.^{46,85,87} For each isotope, a quantized correction curve was fit to the PI simulation data using an inverted Eckart function. The curve fitting was done using the Levenberg–Marquardt algorithm, and the inverted Eckart potential was added to the CM-PMF to obtain the QM-PMF.

Kinetic Isotope Effects. For the primary KIE ($k_{\text{H}}^{\text{H}}/k_{\text{D}}^{\text{H}}$), the *pro-R* hydrogen (the donated hydride, H4N) is substituted with deuterium, whereas the secondary KIE ($k_{\text{H}}^{\text{H}}/k_{\text{D}}^{\text{H}}$) involves the geminal *pro-S* hydrogen (H42N). To evaluate the KIEs, the centroid path-integral simulations were carried out for the light isotopic reaction, and the ratio of the partition functions between two isotopic reactions was determined by free-energy perturbation theory from the light mass into a heavier one.⁴⁶ In the present study, we quantized the donor (C4N_{NADPH}) and acceptor (C6_{H₃folate}) carbons, in addition to the *pro-R* (H4N) and *pro-S* (H42N) hydrogens connected to the donor carbon (in the reactant state). Each quantized particle was represented by 32 beads. We used a bisection sampling technique⁸⁸ in all centroid path-integral simulations, and 10 free-particle configurations were sampled for each of 10 200 classical configurations, yielding a total of 102 000 path-integral sampling steps.

RESULTS AND DISCUSSION

Gas Phase QM Calculations. Optimized Geometries. According to the X-ray crystal structure of synthetically prepared 6-methyl-7,8-dihydropterin-mono-hydrochloride-mono-hydrate (6-Me-H₂pterin·HCl, which is essentially 6-Me-H₃pterin⁺), the heterocyclic ring members form a perfectly planar structure.⁹⁸ However, our gas phase DFT calculations, in particular at the B3LYP/6-31G(2df,p) level, suggest that the framework of the pyrazine ring in the dihydropterins is not completely planar, but

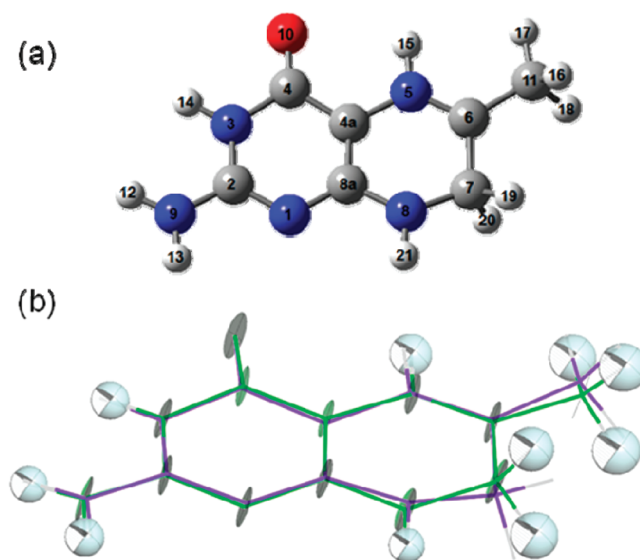


Figure 3. (a) Optimized geometry of 6-Me-H₃pterin⁺ calculated at the B3LYP/6-31G(2df,p) level of theory. (b) Superimposition of the X-ray crystal structure of 6-methyl-7,8-dihydropterin-mono-hydrochloride-mono-hydrate (green) with the computed geometry of 6-Me-H₃pterin⁺ (purple). Isotropic and anisotropic temperature factors corresponding to atoms in the crystal structure are represented by thermal ellipsoids.

rather a pucker is observed as the C7 methylene carbon projects above or below the plane formed by all other ring atoms (Figure 3). The magnitude of deviation from the plane varies with the selection of the method and basis set. MP2/6-311+G(2d,p) optimization supports the nonplanar nature of 6-Me-H₃pterin⁺ predicted by DFT. The discrepancy between experimental results and theory could be due to crystal forces in the experimental structure. Conversely, the observed crystal structure could be the average of two structures puckered in opposite directions. As for the reduced equilibrium structures (6-Me-H₄pterin and 6-Me-H₃pterin⁻), the pyrazine ring adopts a half-chair conformation, where C6 and C7 lie below and above the plane, respectively, in a staggered conformation with respect to each other, and the methyl group attached to C6 takes a pseudo-equatorial position. These findings are in good agreement with ¹H NMR studies on tetrahydropterin derivatives (spin–spin coupling constants measurements).^{99–102} A comparison of the crystal structure of 6-Me-H₃pterin⁺ and the computed geometry at the B3LYP/6-31G(2df,p) level is available in the Supporting Information.

The degree of pyramidalization of the C2-exocyclic amino group can be defined by its torsion angles and tilting from the plane. While in the 6-Me-H₃pterin⁺ structure, this group is nearly coplanar with the 4-oxypyrimidine ring, it exhibits a substantial pyramidalization in the other pterin derivatives examined. As this amino group can be seen as a fragment of a guanidine-like moiety, it is relevant to mention that the solid-state structure of free base guanidine, recently determined by X-ray diffraction,¹⁰³ indicates a nonplanar geometry with pyramidal amino groups, in accordance with earlier *ab initio* calculations.¹⁰⁴ A pyramidal geometry is also exhibited at the N8 position in 6-Me-H₃pterin⁻ (but not in the oxidized form), as well as at the N5 site in 6-Me-H₄pterin.

As for the nicotinamides, it is noteworthy that the *cisoid* conformer was found to be slightly more stable *in vacuo* (with the exception of the unsubstituted neutral molecules, *cis/trans-nic*), while the conformer identified in X-ray crystal structures of many

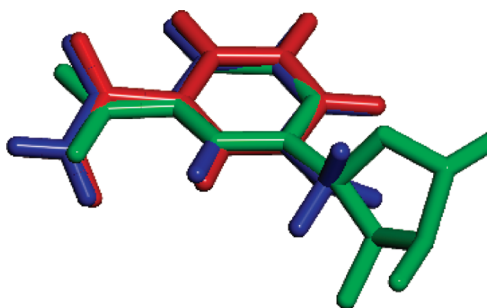


Figure 4. Superimposition of the oxidized nicotinamide portion in three models: segment from the NADP⁺ cofactor (green), coordinates taken from the crystal structure of *E. coli* DHFR ternary complex (PDB ID: 1rx2), *trans*-Me-Hnic⁺ (blue), and *trans*-Hnic⁺ (red) computed with B3LYP/6-31G(2df,p). Note the differences in the orientation of the carboxamide relative to the pyridine ring.

enzyme active sites is the *transoid* form.^{20,67,68} One may attribute the preference of the *transoid* species in the enzyme to specific interactions with the environment (hydrophilic/hydrophobic, polar/nonpolar), which may offset the relatively small energy difference between the two conformers.¹⁰⁵ There are some noticeable differences between the geometries of the *cisoid* and the *transoid* conformers. The degree of distortion of the carboxamide group with respect to the pyridine plane can be expressed in terms of the dihedral angle C2–C3–C7=O8 (Chart 1), as the difference between a “perfect” antiperiplanar angle (180°, for the *transoid* conformer) or synperiplanar angle (0°, for the *cisoid* conformer) and the actual dihedral. Calculations with B3LYP/6-31G(2df,p) suggest that in the *transoid* species, the carboxamide is rotated substantially out-of-plane, with distortions of 18.4–19.7° and 27.5–28.4° for the dihydronicotinamides and their corresponding oxidized counterparts. However, the out-of-plane rotations observed in the *cisoid* conformers are considerably smaller, with values of ca. 2.1° and 13.0–13.2° for H₂nic and Hnic⁺ derivatives, respectively. These observations are in agreement with previous theoretical studies.^{105–108} In the X-ray crystal structure of the ternary complex *ec*DHFR:folate:NADP⁺ (PDB code: 1rx2) used in the QM/MM simulations in this study, the nicotinamide subunit is found in its *transoid* conformer, with out-of-plane distortion of only 6° (Figure 4). Moreover, in most of the *transoid* conformers, there is a considerable pyramidalization of the amide nitrogen. These two geometric features of the carboxamide—the out-of-plane distortion and the *N*-pyramidalization—help to relieve unfavorable steric interactions between the NH₂ group and C2–H bond in the pyridine ring.

The 1,4-dihydropyridine ring skeleton in both unsubstituted and *N*-Me-substituted dihydronicotinamides is nearly planar, and no apparent puckering into a boat conformation is observed. This finding is in accord with X-ray data for some synthetically produced and isolated *N*-substituted nicotinamides, which are essentially planar in the dihydropyridine ring system.^{109,110}

Model Reactions. In Table 1, the calculated energies, enthalpies, and free energies at room temperature are given for the “protonated” (a) and “unprotonated” (b) model reactions involving *N*-methyl substituted nicotinamides. The results are reported separately for each of the four thermodynamic pathways described in the Methodology section (eqs 1a, 1b).

All methods except AM1 predict an exothermic (and exergonic) reaction for the “protonated” model. The reaction energies can be ordered in absolute magnitude as follows: *cisoid* → *transoid* <

cisoid → *cisoid* ≈ *transoid* → *transoid* < *transoid* → *cisoid* (this trend is opposite for the “unprotonated” model). The variation between these pathways stems from the slightly higher energy of the *transoid* conformer of the nicotinamide species relative to its *cisoid* counterpart, either in the reduced or oxidized form. For the density functionals, the relative order of ΔE_{el} (and generally also of ΔH_r and ΔG_r) in absolute values is PBEPBE < M06 < PBE1PBE < BB1K ≈ MPWB1K < B3LYP ≈ B98. The results obtained with CBS-QB3 and G4MP2 are almost identical yet considerably more exothermic than those obtained from DFT (ΔE_{el} being ~3 kcal/mol more negative than in the case of B3LYP, which is closest to the *ab initio* target data).

On the other hand, the “unprotonated” model reaction was found to be highly endothermic (and endergonic) with most of the computed ΔG_r values ranging between 120 and 134 kcal/mol. This emphasizes the crucial role of N5-protonation at the pyrazine ring of pterin in facilitating a thermodynamically feasible hydride transfer from dihydronicotinamide to dihydropterin in the gas phase.

Reaction, Transition, and Product Complexes. The diagram in Figure 5 compares the relative energies of the reactant, transition, and product states (RS, TS, and PS), for the reaction 6-Me-H₃pterin⁺ + *trans*-Me-H₂nic → 6-Me-H₄pterin + *trans*-Me-Hnic⁺, using AM1 and various density functionals.

Focusing on the bimolecular complexes, we define the energy barrier, ΔE^\ddagger , as the energy gap between the TS and RS complexes, and the reaction energy ΔE_r as the difference between the PS and RS complexes. These quantities are poorly predicted by the standard AM1 method, as demonstrated by a high barrier ($\Delta E^\ddagger = 27.8$ kcal/mol) and a slightly endothermic reaction ($\Delta E_r = 1.9$ kcal/mol). The DFT methods suggest a much lower barrier and an exothermic reaction. The gradient corrected functional PBEPBE, which does not contain Hartree–Fock exchange, yields the lowest barrier, $\Delta E^\ddagger = 3.5$ kcal/mol. There is good agreement between the functionals B3LYP, BB1K, MPWB1K, and M06, while B3LYP-D gives a slightly lower barrier. It is interesting to note that the computed gas-phase barriers (e.g., using M06, $\Delta E^\ddagger = 12.2$ kcal/mol) are quite similar to the experimental free-energy barrier in the enzyme, 13.4 kcal/mol.⁵

The effect of dispersion is clearly seen by inspecting the energies of the fully separated molecules relative to the RS/PS complexes. Using the M06 and B3LYP-D functionals, which both account for dispersion interactions, there is a considerable additional stabilization of the complexes. At the B3LYP-D level, this dispersion effect may be estimated directly by comparing with B3LYP, yielding ca. 11 kcal/mol for the RS and PS. We note that the PS complexes presented in Figure 5 correspond to a nearly T-stacked conformation between the pterin and nicotinamide rings, similar to that found in DHFR. For methods that include dispersion, an additional minimum configuration corresponding to planar stacking between the rings is found. This point is discussed further below.

Considering the close agreement between B3LYP and M06 for the current system, we employ B3LYP (and B3LYP-D) as target data for the complexes in the SRP parametrization process. This choice is consistent with the target data for the individual molecules being taken from the G4MP2 approach, which employs B3LYP geometries and thermodynamic data.

SRP Development. The AM1 parameters were reoptimized against high-level QM results, obtained from G4MP2 calculations as well as B3LYP and B3LYP-D calculations with the 6-31+G(d,p) basis set. The properties used as target values

Table 1. Energetics Calculated for the “Protonated” (a) and “Unprotonated” (b) Model Reactions in the Gas Phase^a

method	Part a											
	<i>cisoid</i> → <i>cisoid</i> ^b			<i>cisoid</i> → <i>transoid</i> ^c			<i>transoid</i> → <i>transoid</i> ^d			<i>transoid</i> → <i>cisoid</i> ^e		
	ΔE_{el}	ΔH_r	ΔG_r	ΔE_{el}	ΔH_r	ΔG_r	ΔE_{el}	ΔH_r	ΔG_r	ΔE_{el}	ΔH_r	ΔG_r
AM1		7.4	8.5		9.0	9.8		7.1	7.6		5.5	6.4
B3LYP/6-31+G(d,p)	-7.5	-6.9	-6.6	-5.6	-5.0	-4.6	-7.7	-7.0	-7.0	-9.6	-9.0	-9.0
PBE1PBE/6-31+G(d,p)	-6.7	-6.1	-5.6	-4.7	-4.0	-3.8	-6.9	-6.3	-6.5	-8.8	-8.3	-8.3
PBEPBE/6-31+G(d,p)	-5.3	-4.6	-4.2	-3.5	-2.7	-2.8	-5.2	-4.7	-4.8	-7.0	-6.6	-6.2
B98/6-31+G(d,p)	-7.6	-7.0	-6.5	-5.7	-5.1	-5.5	-7.7	-7.2	-7.8	-9.6	-9.1	-8.7
BB1K/6-31+G(d,p)	-7.0	-6.5	-5.7	-4.9	-4.5	-4.5	-7.3	-6.8	-7.2	-9.4	-8.9	-8.5
MPWB1K/6-31+G(d,p)	-7.1	-6.7	-5.5	-5.0	-4.5	-5.1	-7.4	-6.9	-7.8	-9.5	-9.1	-8.2
M06/6-31+G(d,p)	-6.3	-5.8	-4.8	-4.5	-4.0	-2.2	-6.6	-6.0	-4.8	-8.4	-7.9	-7.4
G4MP2	-10.7	-10.1	-9.5	-8.7	-8.1	-7.6	-10.0	-9.5	-9.0	-12.0	-11.5	-10.9
CBS-QB3	-10.8	-10.1	-9.8	-8.7	-8.0	-7.7	-10.2	-9.6	-9.5	-12.2	-11.7	-11.6

method	Part b											
	<i>cisoid</i> → <i>cisoid</i> ^f			<i>cisoid</i> → <i>transoid</i> ^g			<i>transoid</i> → <i>transoid</i> ^h			<i>transoid</i> → <i>cisoid</i> ⁱ		
	ΔE_{el}	ΔH_r	ΔG_r	ΔE_{el}	ΔH_r	ΔG_r	ΔE_{el}	ΔH_r	ΔG_r	ΔE_{el}	ΔH_r	ΔG_r
AM1		129.6	130.7		131.1	132.0		129.2	129.8		127.7	128.6
B3LYP/6-31+G(d,p)	125.1	124.9	125.1	127.0	126.8	127.1	124.9	124.8	124.7	123.0	122.8	122.7
PBE1PBE/6-31+G(d,p)	127.2	126.8	127.2	129.2	128.8	129.0	127.0	126.5	126.3	125.1	124.5	124.5
PBEPBE/6-31+G(d,p)	121.7	121.1	121.2	123.5	122.9	122.6	121.8	120.9	120.5	120.0	119.1	119.1
B98/6-31+G(d,p)	125.5	124.9	125.4	127.4	126.9	126.3	125.4	124.8	124.0	123.6	122.9	123.1
BB1K/6-31+G(d,p)	131.4	131.0	131.8	133.6	133.0	133.0	131.2	130.7	130.3	129.1	128.6	129.1
MPWB1K/6-31+G(d,p)	131.7	131.2	132.5	133.9	133.3	132.9	131.5	131.0	130.2	129.3	128.8	129.8
M06/6-31+G(d,p)	129.0	128.2	129.0	130.8	130.1	131.6	128.7	128.0	129.0	126.9	126.2	126.4
G4MP2	124.9	124.3	124.7	126.9	126.3	126.5	125.6	125.0	125.1	123.6	123.0	123.2
CBS-QB3	125.2	124.6	124.7	127.2	126.7	126.8	125.8	125.1	125.0	123.7	123.0	122.9

^a ΔE_{el} is the change in electronic energy; ΔH_r and ΔG_r are the respective enthalpy and free energy changes at 298.15 K and 1 atm (including zero-point contributions). ^b 6-Me-H₃pterin⁺ + *cis*-Me-H₂nic ⇌ 6-Me-H₄pterin + *cis*-Me-Hnic⁺. ^c 6-Me-H₃pterin⁺ + *cis*-Me-H₂nic ⇌ 6-Me-H₄pterin + *trans*-Me-Hnic⁺. ^d 6-Me-H₃pterin⁺ + *trans*-Me-H₂nic ⇌ 6-Me-H₄pterin + *trans*-Me-Hnic⁺. ^e 6-Me-H₃pterin⁺ + *trans*-Me-H₂nic ⇌ 6-Me-H₄pterin + *cis*-Me-Hnic⁺. ^f 6-Me-H₂pterin + *cis*-Me-H₂nic ⇌ 6-Me-H₃pterin⁻ + *cis*-Me-Hnic⁺. ^g 6-Me-H₂pterin + *cis*-Me-H₂nic ⇌ 6-Me-H₃pterin⁻ + *trans*-Me-Hnic⁺. ^h 6-Me-H₂pterin + *trans*-Me-H₂nic ⇌ 6-Me-H₃pterin⁻ + *trans*-Me-Hnic⁺. ⁱ 6-Me-H₂pterin + *trans*-Me-H₂nic ⇌ 6-Me-H₃pterin⁻ + *cis*-Me-Hnic⁺.

included heats of formation, dipole moments, Mulliken charges, and vibrational frequencies of the individual species. Furthermore, the reaction energies for the model complexes also served as reference data that contributed to the fitness function of the SRP model (*vide infra*), in order to improve the accuracy of the resulting PES.

Determination of the target heats of formation required some caution, as we are interested in a purely electronic structure description of the potential energy surface, while nuclear classical and quantum effects are added only at a later stage during enzyme simulations. Therefore, any thermal contribution should be excluded. On the other hand, the AM1 energies are originally interpreted in terms of heats of formations.⁵⁸ In the present AM1-SRP reparametrization, we mainly use *ab initio* and DFT energies as target data, and therefore the resulting AM1-SRP energies will mimic electronic energies. This allows the AM1-SRP models to be employed in molecular dynamics simulations where all classical thermal effects are included directly via the propagation of Newton's equations of motion, while all quantum thermal effects are included via path-integral simulations.

Table 2 compares the root-mean-square deviations (RMSD) from the target data for various properties calculated at the

optimized geometries in three parametrization schemes: standard AM1, AM1-SRP, and AM1-SRP(D). The target data for the 14 molecules are the same for the three parametrization schemes. The AM1-SRP parametrization did not involve the heats of formation of the pterin molecules due to the large gap between the corresponding target and AM1 values, an obstacle which was circumvented by minimizing the errors associated with related relative energies (see Table 3). Indeed, AM1-SRP produced the smallest RMSD errors in relative energies among the parametrization schemes, and the quality of the individual heats of formations was considerably improved by ca. 8 kcal/mol when going from standard AM1 to AM1-SRP. There were significant changes in the AM1-SRP(D) parameters compared with standard AM1, in particular in the Gaussian terms. This further reduced the RMSD error in the heats of formation to ca. 1 kcal/mol, including the pterin species (whose heats of formation were incorporated in the training set of AM1-SRP(D)). At the same time, the error in the relative energies was increased only slightly relative to AM1-SRP, making its energetic accuracy comparable to the latter. The RMSD error for the vibrational wavenumbers was reduced by 29 cm⁻¹ and 55 cm⁻¹ in AM1-SRP and AM1-SRP(D), respectively, relative to AM1. On the other hand, the

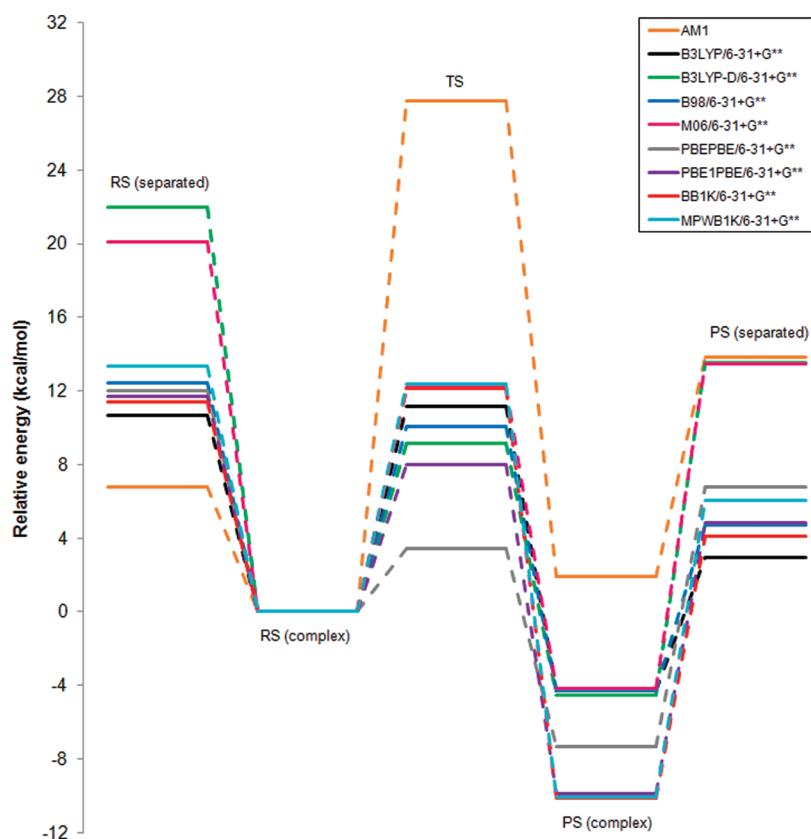


Figure 5. The relative energies (kcal/mol) of the reactant, transition, and product states associated with the reaction $6\text{-Me-H}_3\text{pterin}^+ + \text{trans-Me-H}_2\text{nic} \rightarrow 6\text{-Me-H}_4\text{pterin} + \text{trans-Me-Hnic}^+$. The fully separated reactants and products are represented by RS (separated) and PS (separated), respectively, where RS (complex) and PS (complex) refer to the corresponding bimolecular complexes.

Table 2. RMSD Errors for Properties Calculated at Optimized Geometries Using AM1, AM1-SRP, and AM1-SRP(D), with Respect to Target Values (G4MP2)

properties	AM1	AM1-SRP	AM1-SRP(D)
heats of formation ^a (kcal/mol)	12.1	3.8	1.3
relative energies ^b (kcal/mol)	9.5	1.9	2.2
bond lengths ^c (Å)	0.02	0.03	0.04
bond angles ^c (deg)	1.8	3.4	1.9
vibrational wavenumbers ^d (cm ⁻¹)	108	78	52
Mulliken atomic charges ^e (e)	0.16	0.27	0.20
dipole moments ^e (D)	1.27	1.08	(1.47)

^aThe statistics refer to nicotinamide derivatives only. The target heats of formation for the pterin species were far off the AM1 values and were thus omitted from the calibration set for AM1-SRP [but included in the AM1-SRP(D) parametrization]. ^bThe statistics refer to the reactions described in Table 3. ^cThe statistics refer to all 14 individual species. ^dThe statistics refer to wavenumbers larger than 2500 cm⁻¹ for all 14 individual species. ^eThe statistics refer to eight neutral species out of the 14 individual species. Dipole moments were not included as reference data in the parametrization of AM1-SRP(D), hence the parentheses.

RMSD of the bond lengths, bond angles, and atomic charges calculated with AM1-SRP were slightly increased relative to AM1 by 0.01 Å, 1.6°, and 0.11e, whereas AM1-SRP(D) did somewhat better with 0.02 Å, 0.1°, and 0.04e, respectively. However, their overall quality was considered to be acceptable in light of the

large improvements for the energies and vibrational frequencies. The dipole moments calculated with all three schemes were of similar quality.

Table 3 provides further insight into the quality of selected relative energies obtained with each Hamiltonian. It lists the deviations for various reactions involving four and two species (classes I and II, respectively), with respect to the reference entries. As demonstrated by the total RMSD error for each class, AM1-SRP performs better for reactions of class I, while generally AM1-SRP(D) treats class II transitions more accurately (though to a less pronounced degree).

Having the tailored semiempirical Hamiltonians in hand, some structural and thermodynamic features of the reactant, transition, and product state complexes were analyzed and compared against the target DFT methods (Table 4). The analysis includes geometric relations between the hydride-donating and accepting carbons (C4' and C6, respectively) and the transferring hydride itself, as well as the imaginary vibrational frequency and energetic relations. The three complexes are depicted in Figure 6, comparing geometries that were obtained using DFT and SRP Hamiltonians.

Overall, there is good qualitative agreement between the computed semiempirical structures of the reactant and transition state complexes and the corresponding DFT reference structures, and the improvements of the SRP Hamiltonians over standard AM1 are substantial. In particular, the AM1-SRP Hamiltonian is able to reproduce the imaginary vibrational frequency of the reactive normal mode in the transition state as

Table 3. Signed Errors of Relative Energies (kcal/mol) Calculated with the AM1-SRP and AM1-SRP(D) Parameters, with Respect to the Target Data (G4MP2)^a

class	reaction	signed error (kcal/mol)		
		AM1	AM1-SRP	AM1-SRP(D)
I	6-Me-H ₃ pterin ⁺ + <i>trans</i> -Me-H ₂ nic → 6-Me-H ₄ pterin + <i>trans</i> -Me-Hnic ⁺	17.1	1.0	3.1
	6-Me-H ₃ pterin ⁺ + <i>cis</i> -Me-H ₂ nic → 6-Me-H ₄ pterin + <i>cis</i> -Me-Hnic ⁺	18.2	1.9	4.2
	6-Me-H ₃ pterin ⁺ + <i>trans</i> -H ₂ nic → 6-Me-H ₄ pterin + <i>trans</i> -Hnic ⁺	14.2	-2.3	-1.1
	6-Me-H ₃ pterin ⁺ + <i>cis</i> -H ₂ nic → 6-Me-H ₄ pterin + <i>cis</i> -Hnic ⁺	15.1	-1.4	-0.2
	6-Me-H ₂ pterin + <i>trans</i> -Me-H ₂ nic → 6-Me-H ₃ pterin ⁻ + <i>trans</i> -Me-Hnic ⁺	3.8	0.6	2.0
	6-Me-H ₂ pterin + <i>cis</i> -Me-H ₂ nic → 6-Me-H ₃ pterin ⁻ + <i>cis</i> -Me-Hnic ⁺	4.8	1.6	3.1
	6-Me-H ₂ pterin + <i>trans</i> -H ₂ nic → 6-Me-H ₃ pterin ⁻ + <i>trans</i> -Hnic ⁺	0.9	-2.7	-2.2
	6-Me-H ₂ pterin + <i>cis</i> -H ₂ nic → 6-Me-H ₃ pterin ⁻ + <i>cis</i> -Hnic ⁺	1.8	-1.7	-1.3
RMSD		11.7	1.8	2.5
II	6-Me-H ₃ pterin ⁺ $\xrightarrow{\text{H}^-}$ 6-Me-H ₄ pterin	15.7	1.6	1.3
	6-Me-H ₂ pterin $\xrightarrow{\text{H}^-}$ 6-Me-H ₃ pterin ⁻	2.4	1.3	0.2
	<i>trans</i> -Me-H ₂ nic $\xrightarrow[\text{H}^-]{\text{H}^-}$ <i>trans</i> -Me-Hnic ⁺	1.4	-0.6	1.8
	<i>cis</i> -Me-H ₂ nic $\xrightarrow[\text{H}^-]{\text{H}^-}$ <i>cis</i> -Me-Hnic ⁺	2.5	0.3	3.0
	<i>trans</i> -H ₂ nic $\xrightarrow[\text{H}^-]{\text{H}^-}$ <i>trans</i> -Hnic ⁺	-1.5	-3.9	-2.4
	<i>cis</i> -H ₂ nic $\xrightarrow[\text{H}^-]{\text{H}^-}$ <i>cis</i> -Hnic ⁺	-0.6	-3.0	-1.5
	<i>cis</i> -nic → <i>trans</i> -nic	0.5	0.3	0.6
RMSD		6.1	2.1	1.8

^a Classes I and II assemble relative energies involving four and two species, respectively.

predicted by B3LYP/6-31+G(d,p) calculations (i.e., -880 cm^{-1} vs -838 cm^{-1}). On the other hand, the calculated imaginary frequency of the TS complex with AM1-SRP(D), -960 cm^{-1} , resembles more that predicted by M06, -953 cm^{-1} , than the corresponding B3LYP-D value, -716 cm^{-1} .

The energetic profiles obtained at different theoretical levels are presented in Figure 7. As expected, the AM1-SRP complexation energies are fairly close to those of B3LYP (-9.3 and -10.7 kcal/mol for RS, -3.0 and -7.2 kcal/mol for PS, respectively), while the AM1-SRP(D) results are similar to those of B3LYP-D (-22.8 and -22.0 kcal/mol for RS, -25.2 and -27.0 kcal/mol for PS; the counterpoise corrections for basis set superposition error (BSSE) with B3LYP-D are 1.5 and 2.1 kcal/mol for RS and PS, respectively). The barrier height obtained with AM1-SRP is 10.7 kcal/mol, comparable to 11.2 kcal/mol with B3LYP, while that obtained with AM1-SRP(D) is 13.6 kcal/mol, which is somewhat higher than the target value 9.2 kcal/mol obtained with B3LYP-D. In comparison, the M06 barrier height is 12.2 kcal/mol. An in-depth analysis of the RS and PS complexes will be presented below.

A detailed inspection of the geometry optimizations of the RS and PS complexes reveals that the PES in these regions is very flat with several plausible minima, particularly in the PS region. Using B3LYP and the AM1-SRP Hamiltonian, we could attain reasonable geometries of the RS and PS complexes via the IRC path from the TS only down to a certain point away from the TS in either direction (these structures are presented in Figure 6).

Full geometry optimization of the RS and PS complexes in the gas phase yielded highly distorted minimum structures, with no stacking, which do not resemble the configuration in the DHFR active site. As the two molecular segments in these complexes are weakly bound in the van der Waals region ($>3.5 \text{ \AA}$), it is necessary to include dispersion interactions (which are missing in the AM1 formalism) to accurately reproduce these geometries. Furthermore, the target complexes for AM1-SRP were derived from B3LYP calculations which do not include much dispersion. Indeed, the introduction of dispersion by means of B3LYP-D enabled us to obtain plausible minimum RS and PS geometries.

B3LYP-D and M06 optimizations of the PS both reveal two possible minimum configurations, with regard to the orientation of the nicotinamide ring toward the pterin fragment. (a) First is a "T-stacked" configuration, where the nicotinamide ring lies in a quasi-perpendicular plane with respect to the pterin ring. With this orientation, the complexation energy obtained with B3LYP-D is -18.1 kcal/mol , while M06 affords a related minimum geometry with a complexation energy of -17.6 kcal/mol . (b) Next is a fully "stacked" configuration, in which the rings are stacked one on top of the other, which corresponds to the global minimum of the PES of the PS complex (Figures 6 and 7). The B3LYP-D structure (complexation energy: -27.0 kcal/mol) is almost identical to the fully stacked complex obtained with M06 (-23.3 kcal/mol) and resembles that obtained with AM1-SRP(D) (-25.2 kcal/mol). With AM1-SRP(D), we could not find a partially T-stacked configuration, suggesting that this

Table 4. Geometric and Thermodynamic Properties of the Reactant, Transition and Product State Complexes in the Gas Phase, Calculated with Different Hamiltonians^a

complex	method	R (Å)			\angle C4'–H–C6 (deg)	IFreq (cm ⁻¹)	ΔE (kcal/mol)
		C4'–H	C6–H	C4'–C6			
reactant state	AM1 ^b	1.128	2.640	3.636	146.61		0.0
	B3LYP/6-31+G(d,p) ^b	1.110	2.996	3.863	135.95		0.0
	AM1-SRP ^b	1.110	2.551	3.593	155.91		0.0
	B3LYP-D/6-31+G(d,p) ^c	1.103	2.751	3.500	124.85		0.0
	AM1-SRP(D) ^c	1.087	2.283	3.170	137.49		0.0
	M06 ^c	1.104	3.301	3.944	118.15		0.0
transition state	AM1	1.411	1.363	2.763	169.81	-1385	27.8
	B3LYP/6-31+G(d,p)	1.309	1.409	2.715	174.61	-838	11.2
	AM1-SRP	1.317	1.348	2.633	162.43	-880	10.7
	B3LYP-D/6-31+G(d,p)	1.280	1.396	2.664	169.09	-716	9.2
	AM1-SRP(D)	1.308	1.322	2.602	163.18	-960	13.6
	M06	1.313	1.362	2.667	171.01	-953	12.2
product state	AM1 ^b	2.512	1.136	3.507	145.37		1.9
	B3LYP/6-31+G(d,p) ^b	2.533	1.110	3.613	163.96		-4.3
	AM1-SRP ^b	2.514	1.116	3.503	147.07		-2.8
	B3LYP-D/6-31+G(d,p) ^{c,d}	2.679	1.110	3.632	143.52		-4.6
		(5.069)	(1.100)	(5.394)	(101.32)		(-13.5)
	AM1-SRP(D) ^c	4.248	1.092	4.748	110.82		-9.3
M06 ^{c,d}	2.499	1.107	3.443	142.35		-4.2	
	(5.034)	(1.107)	(5.706)	(107.44)		(-9.8)	

^a C4' is the donor carbon in the nicotinamide subunit, C6 is the acceptor carbon in the pterin subunit, and H denotes the transferring hydride. IFreq is the imaginary frequency, and ΔE is the relative energy with respect to the reactant complex. ^b The final structure was obtained using IRC calculations. ^c The final structure was obtained using geometry optimization. ^d Values without parentheses refer to the local minimum structure with a "T-stacked" configuration. Values in parentheses refer to the fully stacked configuration, which corresponds to the global minimum on the potential energy surface.

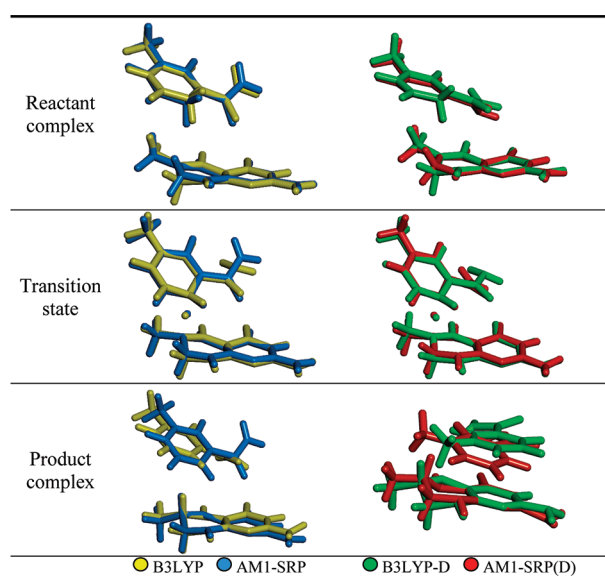


Figure 6. Structures of the reactant, transition, and product state complexes in the gas phase. The superimposed geometries were obtained by calculations at a DFT and a related semiempirical level. Structures are shown for two pairs of methods: B3LYP/6-31+G(d,p) and AM1-SRP and B3LYP-D/6-31+G(d,p) and AM1-SRP(D). The product state complexes represent the T-stacked configuration for B3LYP and AM1-SRP optimized structures and the fully stacked configuration for the B3LYP-D and AM1-SRP(D) optimized structures.

method might "over-stack" the PS complex. The greater stability of the fully stacked complexes may be attributed to the presence of dispersion interactions, hydrogen bonding interactions within the bimolecular complex, and close contacts between the π systems in the nicotinamide and pterin rings.

The gradient norm of the product complex during AM1-SRP(D) geometry optimization is plotted in Figure S1 (Supporting Information). The gradient is greatly reduced in the first steps. The PES then becomes quite flat and shallow. Thus, the structure of the product complex seems to be highly flexible, and many arrangements of the two molecular segments have relatively small gradients. The use of less stringent convergence criteria in the AM1-SRP(D) geometry optimization would have led to termination after about 100 cycles, at a structure close to the T-stacked configuration obtained with B3LYP-D (Figure S1).

QM/MM Interactions. To investigate the ability of the AM1-SRP Hamiltonians to accurately model QM/MM interactions, we computed the complexation energies between selected QM moieties (6-Me-H₃pterin⁺, *trans*-Me-H₂nic, 6-Me-H₄pterin, and *trans*-Me-Hnic⁺) and a TIP3P water molecule. For each of these QM moieties, a single water molecule was placed at different hydrogen bonding positions around the molecule, for a total of 14 QM/MM complexes (Table S8a, Supporting Information). The QM/MM interaction energies were computed using AM1, AM1-SRP, or AM1-SRP(D) and a TIP3P water molecule. These interaction energies were compared with data from M06, B3LYP-D, B3LYP, and HF, all in conjunction with the

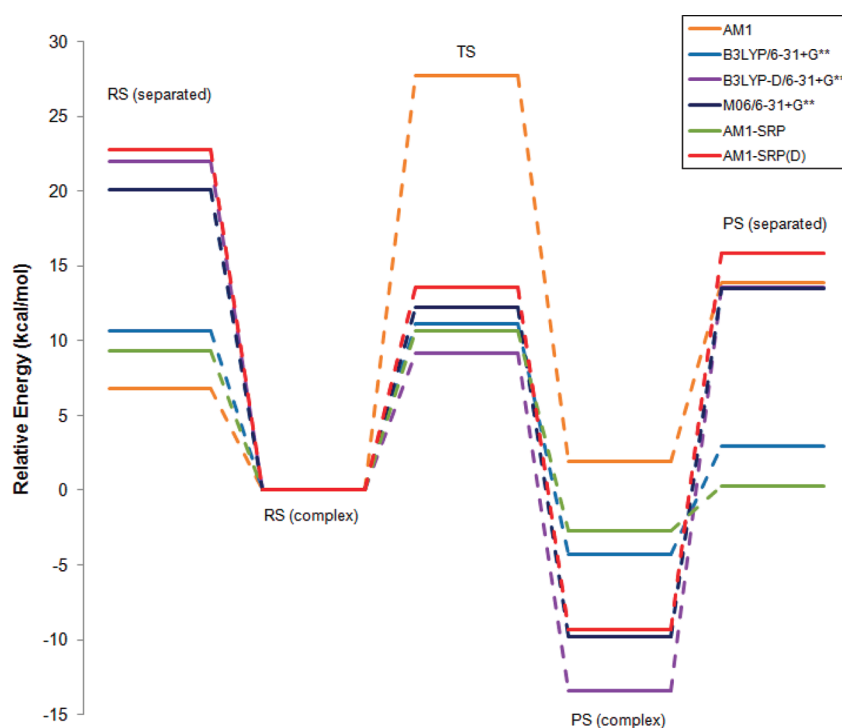


Figure 7. The relative energies of the reactant, transition and product states associated with the reaction $6\text{-Me-H}_3\text{pterin}^+ + \text{trans-Me-H}_2\text{nic} \rightarrow 6\text{-Me-H}_4\text{pterin} + \text{trans-Me-Hnic}^+$, for standard AM1, B3LYP/6-31+G**, B3LYP-D/6-31+G**, M06/6-31+G**, AM1-SRP, and AM1-SRP(D). The sums of the energies of the individual subunits for the reactants and products are represented by RS (separated) and PS (separated), respectively, where RS (complex) and PS (complex) refer to the corresponding bimolecular complexes.

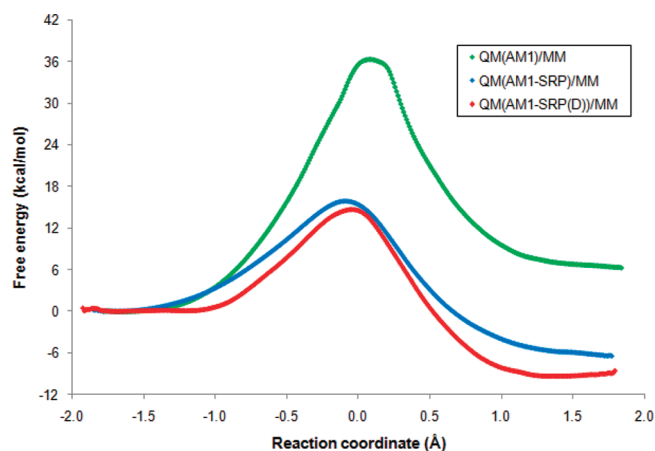


Figure 8. Computed classical potentials of mean force for the hydride transfer reaction catalyzed by *E. coli* DHFR at 298 K, obtained with three different QM/MM schemes distinguished by the semiempirical treatment of the QM region: AM1/MM (green), AM1-SRP/MM (blue), and AM1-SRP(D)/MM (red). The reaction coordinate is defined as the difference between the distances of the transferring hydride and the donor and acceptor carbon atoms.

6-31+G(d,p) basis set. BSSE corrections were included for the DFT methods but not for HF. The different semiempirical QM/MM methods give similar interaction energies, indicating that the SRP optimization process did not introduce artificial polarization of the molecules (Table S8b and Figure S2, Supporting Information). The QM/MM methods predict slightly weaker complexes than M06 and B3LYP-D (RMSDs of 2.6–4.4 kcal/mol),

while they are in good agreement with B3LYP and HF (RMSDs between 1.2 and 2.4 kcal/mol).

Enzyme Simulations. *Classical Potential of Mean Force.* Figure 8 shows the classical mechanical PMF (CM-PMF) for the hydride transfer reaction in DHFR obtained from free energy MD simulations, using the AM1, AM1-SRP, and AM1-SRP(D) QM/MM Hamiltonians. We note that these data are not directly comparable to experimental results as NQEs are not yet included at this stage. The AM1/MM Hamiltonian substantially overestimates the free energy barrier ($\Delta G^\ddagger = 36.3$ kcal/mol) and predicts an endothermic reaction ($\Delta G^r = 6.3$ kcal/mol). The transition state (the free energy bottleneck of the PMF) for the QM(AM1-SRP)/MM Hamiltonian is located at $\zeta^\ddagger = -0.089$ Å, whereas that of the QM(AM1-SRP(D))/MM Hamiltonian is placed at $\zeta^\ddagger = -0.045$ Å. Both are relatively late compared to that reported by Gao and co-workers in recent QM/MM studies ($\zeta^\ddagger = -0.145$ Å) on *ec*DHFR, which utilized the AM1 Hamiltonian with SVB correction (*vide supra*).^{7,8} We further note that the end points of the AM1-SRP and AM1-SRP(D) QM/MM free energy profiles are smooth and show no artificial free-energy increase. The classical free energies of reaction and activation extracted from the QM(AM1-SRP)/MM-based free energy profile are -6.4 and 15.9 kcal/mol. The CM-PMF obtained by the enhanced Hamiltonian, QM(AM1-SRP(D))/MM, predicts a classical mechanical free energy barrier which is narrower and slightly lower ($\Delta G^\ddagger = 14.7$ kcal/mol), and the reaction is more exergonic ($\Delta G^r = -9.3$ kcal/mol) by 3 kcal/mol. The considerable stabilization of the product state by ca. 5 kcal/mol compared with the experimental data is possibly a manifestation of the “over-stacking” phenomenon already encountered with the AM1-SRP(D) in the gas phase (*vide supra*). Finally, the differences

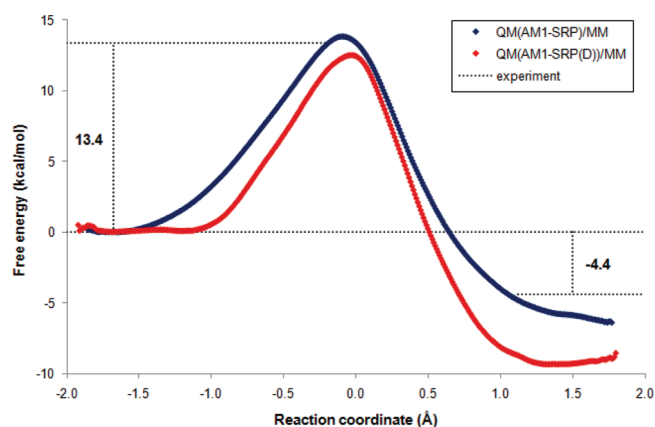


Figure 9. The quantum mechanical potentials of mean force for the hydride transfer reaction in *E. coli* DHFR. The centroid coordinates are used in path-integral simulations. Each quantized particle was represented by 32 beads. The experimental free energies of reaction and activation are denoted near the dashed lines.

in the free energy barrier width among the three QM/MM Hamiltonians are consistent with the trends in the imaginary vibrational frequency of the transition state in the gas phase obtained from the corresponding semiempirical schemes. According to Table 4, the imaginary wavenumbers in absolute values are ordered as follows: AM1 > AM1-SRP(D) > AM1-SRP. Indeed, the QM(AM1)/MM free energy profile has the narrowest barrier, that of QM(AM1-SRP(D))/MM is wider, and the QM(AM1-SRP)/MM Hamiltonian yields a PMF with the widest barrier.

Quantum Potential of Mean Force. The QM-PMF is obtained from Feynman path-integral calculations,^{46,86,88} in which the centroid positions of the discrete paths of quantized particles are used to specify the reaction coordinate.^{46,89,90} The “quantum” free energy profiles displayed in Figure 9 describe the hydride transfer reaction with the two SRP QM/MM Hamiltonians. Using QM(AM1-SRP)/MM, the inclusion of NQE in the simulations⁴⁶ lowers the computed free energies of activation for the hydride and deuteride transfer by 2.1 and 1.4 kcal/mol, respectively, relative to the “classical” free energy barrier. The resulting quantum free energies of reaction and activation for the hydride transfer, -6.4 and 13.8 kcal/mol, are in good accord with the corresponding experimental results (-4.4 and 13.4 kcal/mol).⁵ The quantum corrections for the AM1-SRP(D)/MM CM-PMF are very similar, 2.2 and 1.5 kcal/mol for hydride and deuteride, so that the predicted free energy barrier (12.5 kcal/mol) is in close agreement with experimental results. The free-energy results demonstrate that the present QM/MM and path integral methods can provide an adequate description of the hydride transfer reaction in DHFR.

Kinetic Isotope Effects. The computed primary ($k_{\text{H}}^{\text{H}}/k_{\text{D}}^{\text{H}}$) and secondary ($k_{\text{H}}^{\text{H}}/k_{\text{H}}^{\text{D}}$) KIEs for the hydride transfer reaction in DHFR at 298 K are 3.51 ± 0.14 and 1.18 ± 0.06 with QM(AM1-SRP)/MM and 3.49 ± 0.16 and 1.11 ± 0.04 with QM(AM1-SRP(D))/MM, respectively. These values are in good agreement with the experimental intrinsic KIEs measured by Kohen and co-workers ($k_{\text{H}}^{\text{H}}/k_{\text{D}}^{\text{H}} = 3.55 \pm 0.17$; $k_{\text{H}}^{\text{H}}/k_{\text{H}}^{\text{D}} = 1.13 \pm 0.01$),⁷² providing additional evidence for the accuracy of our computational treatments. These KIEs are also in good agreement with various QM/MM calculations (e.g., refs 7 (2° KIEs = 1.13) and 39 (1° KIEs = 3.4 ± 0.6)), but here the same method is within experimental error for both 1° and 2° KIEs.

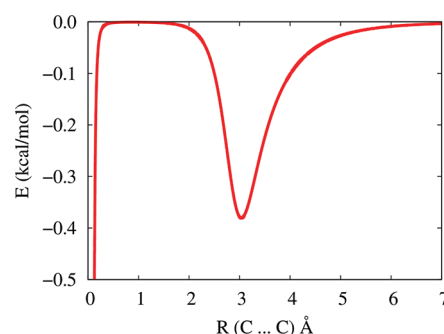


Figure 10. The dispersion energy between two carbon atoms calculated using Grimme's formula.

SUMMARY

In the current work, we presented extensive benchmark calculations for several model reactions in the gas phase that are relevant to the DHFR catalyzed hydride transfer. We employed G4MP2 and CBS-QB3 ab initio calculations as well as numerous density functional methods. Using these results as target data, we developed two specific reaction parameter (SRP) Hamiltonians by reparametrization of the semiempirical AM1 method. The first generation SRP Hamiltonian does not account for dispersion, while the second generation SRP includes dispersion implicitly via the AM1 core-repulsion functions. These SRP semiempirical Hamiltonians were subsequently used in hybrid quantum mechanics/molecular mechanics simulations of the DHFR catalyzed reaction. The classical PMFs were computed using the standard AM1 method as well as the AM1-SRP and AM1-SRP(D) models. Nuclear quantum effects were included using a Feynman path-integral method. Finally, kinetic isotope effects were computed using a mass-perturbation-based path-integral approach. The quantum PMFs predict free energy barriers and reaction free energies in good agreement with available experimental kinetic data.

We conclude that the resulting PESs yield accuracies comparable to those obtained at the G4MP2 and DFT levels, with a computational cost that is several orders of magnitude less. This will allow us to perform long MD simulations of the solvated enzyme, while providing a realistic description of the kinetics and thermodynamic properties in the DHFR catalyzed reaction.

APPENDIX A

The dispersion energies were estimated using the formula introduced by Grimme for density functional methods.

$$E_{\text{disp}}(r_{ij}) = -s_6 \sum_{i=1}^N \sum_{j=i+1}^{N-1} \frac{C_{6,ij}}{r_{ij}^6} f_d(r_{ij})$$

$$f_d(r_{ij}) = \frac{1}{1 + \exp\{-d[(r_{ij}/r_0) - 1]\}}$$

where r_{ij} is the distance between atoms i and j . s_6 is a global scaling factor for the dispersion energy, with numerical values ranging from 0.75 to 1.2 for different density functionals. Herein, we chose the value 1.0 for simplicity. d is a parameter for damping function f_d . It was chosen to be 20.0 in line with the value used in DFT-D2. The atomic C_6 coefficients and van der Waals radii r_0 were taken directly from the original publication. Figure 10 plots the calculated dispersion energy for two carbon atoms.

The sum of the standard AM1 core repulsion function and the dispersion energy correction was scanned for each atom pair. A total of 396 points for the carbon, nitrogen, and oxygen pairs and 346 points for the hydrogen pair were collected from a distance of 0.1 Å with a step size of 0.02 Å. Finally, the parameters in the Gaussian terms were fitted to these data points using a nonlinear least-squares procedure implemented in gnuplot with the standard AM1 parameters as an initial guess. The convergence criterion was 1.0×10^{-6} .

■ ASSOCIATED CONTENT

S Supporting Information. Comprehensive list of the gas phase molecular models; summary of recent computational studies on the DHFR-catalyzed hydride transfer reaction; vibrational frequency scale factors used for gas phase thermochemistry calculations; details of the development of AM1-SRP, calibration, and related target values; selected gas phase vibrational frequencies; list of the semiempirical parameters of AM1-SRP and AM1-SRP(D); coordinates of geometries calculated at the target levels; comparison between the crystal structure and computed geometry of 6-Me-H₃pterin⁺; ligand–water QM/MM complexation energies; and further details with regard to the MD simulations of the solvated enzyme. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: majort@biu.ac.il.

■ ACKNOWLEDGMENT

This work has been supported by the Israel Science Foundation and the United States–Israel Binational Science Foundation (Grant # 2007256).

■ REFERENCES

- (1) Schnell, J. R.; Dyson, H. J.; Wright, P. E. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 119–140.
- (2) Huennekens, F. In search of dihydrofolate reductase. *Protein Sci.* **1996**, *5*, 1201.
- (3) Frey, P. A.; Hegeman, A. D. *Enzymatic reaction mechanisms*; Oxford University Press: New York, 2007.
- (4) Castillo, R.; Andrés, J.; Moliner, V. Catalytic Mechanism of Dihydrofolate Reductase Enzyme. A Combined Quantum-Mechanical/Molecular-Mechanical Characterization of Transition State Structure for the Hydride Transfer Step. *J. Am. Chem. Soc.* **1999**, *121*, 12140–12147.
- (5) Fierke, C. A.; Johnson, K. A.; Benkovic, S. J. Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase from *Escherichia coli*. *Biochemistry* **1987**, *26*, 4085–4092.
- (6) Morrison, J. F.; Stone, S. R. Mechanism of the reaction catalyzed by dihydrofolate reductase from *Escherichia coli*: pH and deuterium isotope effects with NADPH as the variable substrate. *Biochemistry* **1988**, *27*, 5499–5506.
- (7) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. Reaction-Path Energetics and Kinetics of the Hydride Transfer Reaction Catalyzed by Dihydrofolate Reductase. *Biochemistry* **2003**, *42*, 13558–13575.
- (8) Pu, J.; Ma, S.; Garcia-Viloca, M.; Gao, J.; Truhlar, D. G.; Kohen, A. Nonperfect Synchronization of Reaction Center Rehybridization in the Transition State of the Hydride Transfer Catalyzed by Dihydrofolate Reductase. *J. Am. Chem. Soc.* **2005**, *127*, 14879–14886.

- (9) Maharaj, G.; Selinsky, B.; Appleman, J.; Perlman, M.; London, R.; Blakley, R. Dissociation constants for dihydrofolic acid and dihydrobiopterin and implications for mechanistic models for dihydrofolate reductase. *Biochemistry* **1990**, *29*, 4554.
- (10) Chen, Y. Q.; Kraut, J.; Blakley, R. L.; Callender, R. Determination by Raman Spectroscopy of the pKa of N5 of Dihydrofolate Bound to Dihydrofolate Reductase: Mechanistic Implications. *Biochemistry* **1994**, *33*, 7021–7026.
- (11) Deng, H.; Callender, R. Structure of Dihydrofolate When Bound to Dihydrofolate Reductase. *J. Am. Chem. Soc.* **1998**, *120*, 7730–7737.
- (12) Deng, H.; Callender, R.; Howell, E. Vibrational Structure of Dihydrofolate Bound to R67 Dihydrofolate Reductase. *J. Biol. Chem.* **2001**, *276*, 48956–48960.
- (13) Howell, E.; Villafranca, J.; Warren, M.; Oatley, S.; Kraut, J. Functional role of aspartic acid-27 in dihydrofolate reductase revealed by mutagenesis. *Science* **1986**, *231*, 1123–1128.
- (14) David, C. L.; Howell, E. E.; Farnum, M. F.; Villafranca, J. E.; Oatley, S. J.; Kraut, J. Structure and function of alternative proton-relay mutants of dihydrofolate reductase. *Biochemistry* **1992**, *31*, 9813–9822.
- (15) Chen, Y.; Kraut, J.; Callender, R. pH-dependent conformational changes in *Escherichia coli* dihydrofolate reductase revealed by Raman difference spectroscopy. *Biophys. J.* **1997**, *72*, 936–941.
- (16) Rod, T. H.; Brooks, C. L., III. How Dihydrofolate Reductase Facilitates Protonation of Dihydrofolate. *J. Am. Chem. Soc.* **2003**, *125*, 8718–8719.
- (17) Bolin, J.; Filman, D.; Matthews, D.; Hamlin, R.; Kraut, J. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding of methotrexate. *J. Biol. Chem.* **1982**, *257*, 13650.
- (18) Bystroff, C.; Oatley, S.; Kraut, J. Crystal structures of *Escherichia coli* dihydrofolate reductase: The NADP⁺ holoenzyme and the folate·NADP⁺ ternary complex. Substrate binding and a model for the transition state. *Biochemistry* **1990**, *29*, 3263–3277.
- (19) Bystroff, C.; Kraut, J. Crystal structure of unliganded *Escherichia coli* dihydrofolate reductase. Ligand-induced conformational changes and cooperativity in binding. *Biochemistry* **1991**, *30*, 2227–2239.
- (20) Sawaya, M.; Kraut, J. Loop and Subdomain Movements in the Mechanism of *Escherichia coli* Dihydrofolate Reductase: Crystallographic Evidence. *Biochemistry* **1997**, *36*, 586–603.
- (21) Donkersloot, M. C. A.; Buck, H. M. The hydride-donation reaction of reduced nicotinamide adenine dinucleotide. 2. MINDO/3 and STO-3G calculations on the role of the carbamoyl group in enzymic reactions. *J. Am. Chem. Soc.* **1981**, *103*, 6554–6558.
- (22) Cummins, P. L.; Gready, J. E. Mechanistic aspects of biological redox reactions involving NADH 2: A combined semiempirical and *ab initio* study of hydride-ion transfer between the NADH analogue, 1-methyl-dihydro nicotinamide, and folate and dihydrofolate analogue substrates of dihydrofolate reductase. *J. Comput. Chem.* **1990**, *11*, 791–804.
- (23) Andrés, J.; Safont, V. S.; Martins, J. B. L.; Beltrán, A.; Moliner, V. AM1 and PM3 transition structure for the hydride transfer. A model of reaction catalyzed by dihydrofolate reductase. *THEOCHEM* **1995**, *330*, 411–416.
- (24) Andrés, J.; Moliner, V.; Safont, V. S.; Domingo, L. R.; Picher, M. T.; Krechl, J. On Transition Structures for Hydride Transfer Step: A Theoretical Study of the Reaction Catalyzed by Dihydrofolate Reductase Enzyme. *Bioorg. Chem.* **1996**, *24*, 10–18.
- (25) Cummins, P. L.; Ramnarayan, K.; Singh, U. C.; Gready, J. E. Molecular dynamics/free energy perturbation study on the relative affinities of the binding of reduced and oxidized NADP to dihydrofolate reductase. *J. Am. Chem. Soc.* **1991**, *113*, 8247–8256.
- (26) Cummins, P. L.; Gready, J. E. Molecular dynamics and free energy perturbation study of hydride-ion transfer step in dihydrofolate reductase using combined quantum and molecular mechanical model. *J. Comput. Chem.* **1998**, *19*, 977–988.
- (27) Ferrer, S.; Silla, E.; Tuñón, I.; Martí, S.; Moliner, V. Catalytic Mechanism of Dihydrofolate Reductase Enzyme. A Combined Quantum-Mechanical/Molecular-Mechanical Characterization of the N5 Protonation Step. *J. Phys. Chem. B* **2003**, *107*, 14036–14041.

- (28) Cummins, P. L.; Greatbanks, S. P.; Rendell, A. P.; Gready, J. E. Computational Methods for the Study of Enzymic Reaction Mechanisms. I. Application to the Hydride Transfer Step in the Catalysis of Dihydrofolate Reductase. *J. Phys. Chem. B* **2002**, *106*, 9934–9944.
- (29) Thorpe, I. F.; Brooks, C. L. Barriers to Hydride Transfer in Wild Type and Mutant Dihydrofolate Reductase from *E. coli*. *J. Phys. Chem. B* **2003**, *107*, 14042–14051.
- (30) Thorpe, I. F.; Brooks, C. L. Conformational Substates Modulate Hydride Transfer in Dihydrofolate Reductase. *J. Am. Chem. Soc.* **2005**, *127*, 12997–13006.
- (31) Ranganathan, S.; Gready, J. E. Hybrid Quantum and Molecular Mechanical (QM/MM) Studies on the Pyruvate to L-Lactate Interconversion in L-Lactate Dehydrogenase. *J. Phys. Chem. B* **1997**, *101*, 5614–5618.
- (32) Turner, A.; Moliner, V.; Williams, I. Transition-state structural refinement with GRACE and CHARMM: Flexible QM/MM modelling for lactate dehydrogenase. *Phys. Chem. Chem. Phys.* **1999**, *1*, 1323–1331.
- (33) Faulder, P. F.; Tresadern, G.; Chohan, K. K.; Scrutton, N. S.; Sutcliffe, M. J.; Hillier, I. H.; Burton, N. A. QM/MM Studies Show Substantial Tunneling for the Hydrogen-Transfer Reaction in Methylamine Dehydrogenase. *J. Am. Chem. Soc.* **2001**, *123*, 8604–8605.
- (34) Ferrer, S.; Ruiz-Pernía, J. J.; Tuñón, I.; Moliner, V.; Garcia-Viloca, M.; González-Lafont, A.; Lluch, J. M. A QM/MM Exploration of the Potential Energy Surface of Pyruvate to Lactate Transformation Catalyzed by LDH. Improving the Accuracy of Semiempirical Descriptions. *J. Chem. Theory Comput.* **2005**, *1*, 750–761.
- (35) Devi-Kesavan, L. S.; Garcia-Viloca, M.; Gao, J. Semiempirical QM/MM potential with simple valence bond (SVB) for enzyme reactions. Application to the nucleophilic addition reaction in haloalkane dehalogenase. *Theor. Chem. Acc.* **2003**, *109*, 133–139.
- (36) Truhlar, D. G.; Gao, J.; Alhambra, C.; Garcia-Viloca, M.; Corchado, J.; Sánchez, M. L.; Villà, J. The Incorporation of Quantum Effects in Enzyme Kinetics Modeling. *Acc. Chem. Res.* **2002**, *35*, 341–349.
- (37) Warshel, A.; Weiss, R. M. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J. Am. Chem. Soc.* **1980**, *102*, 6218–6226.
- (38) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes*; Wiley: New York, 1991.
- (39) Agarwal, P. K.; Billeter, S. R.; Hammes-Schiffer, S. Nuclear quantum effects and enzyme dynamics in dihydrofolate reductase catalysis. *J. Phys. Chem. B* **2002**, *106*, 3283–3293.
- (40) Wong, K. F.; Selzer, T.; Benkovic, S. J.; Hammes-Schiffer, S. Chemical theory and computation special feature: impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 6807–6812.
- (41) Liu, H.; Warshel, A. The Catalytic Effect of Dihydrofolate Reductase and Its Mutants Is Determined by Reorganization Energies. *Biochemistry* **2007**, *46*, 6011–6025.
- (42) Hwang, J. K.; Warshel, A. A quantized classical path approach for calculations of quantum mechanical rate constants. *J. Phys. Chem.* **1993**, *97*, 10053–10058.
- (43) Rossi, I.; Truhlar, D. G. Parameterization of NDDO wavefunctions using genetic algorithms. An evolutionary approach to parameterizing potential energy surfaces and direct dynamics calculations for organic reactions. *Chem. Phys. Lett.* **1995**, *233*, 231–236.
- (44) Major, D. T.; York, D. M.; Gao, J. L. Solvent polarization and kinetic isotope effects in nitroethane deprotonation and implications to the nitroalkane oxidase reaction. *J. Am. Chem. Soc.* **2005**, *127*, 16374–16375.
- (45) Gao, J. L.; Wong, K. Y.; Major, D. T. Combined QM/MM and path integral simulations of kinetic isotope effects in the proton transfer reaction between nitroethane and acetate ion in water. *J. Comput. Chem.* **2008**, *29*, 514–522.
- (46) Major, D. T.; Gao, J. L. An integrated path integral and free-energy perturbation-umbrella sampling method for computing kinetic isotope effects of chemical reactions in solution and in enzymes. *J. Chem. Theory Comput.* **2007**, *3*, 949–960.
- (47) Major, D. T.; Gao, J. L. A combined quantum mechanical and molecular mechanical study of the reaction mechanism and alpha-amino acidity in alanine racemase. *J. Am. Chem. Soc.* **2006**, *128*, 16345–16357.
- (48) Major, D. T.; Heroux, A.; Orville, A. M.; Valley, M. P.; Fitzpatrick, P. F.; Gao, J. Differential quantum tunneling contributions in nitroalkane oxidase catalyzed and the uncatalyzed proton transfer reaction. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 20734–20739.
- (49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2009.
- (50) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (51) Adamo, C.; Cossi, M.; Barone, V. An accurate density functional method for the study of magnetic properties: the PBE0 model. *THEOCHEM* **1999**, *493*, 145–157.
- (52) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (53) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785.
- (54) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *Ab Initio* Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (55) Schmider, H. L.; Becke, A. D. Optimized density functionals from the extended G2 test set. *J. Chem. Phys.* **1998**, *108*, 9624–9631.
- (56) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Development and Assessment of a New Hybrid Density Functional Model for Thermochemical Kinetics. *J. Phys. Chem. A* **2004**, *108*, 2715–2719.
- (57) Zhao, Y.; Truhlar, D. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (58) Dewar, M. J. S.; Ziegler, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (59) Montgomery, J. J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A complete basis set model chemistry. VII. Use of the minimum population localization method. *J. Chem. Phys.* **2000**, *112*, 6532–6542.
- (60) Montgomery, J. J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A complete basis set model chemistry. VI. Use of density functional geometries and frequencies. *J. Chem. Phys.* **1999**, *110*, 2822–2827.
- (61) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108–12.
- (62) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **2007**, *127*, 124105.
- (63) Armarego, W. L. F.; Waring, P.; Williams, J. W. Absolute configuration of 6-methyl-5,6,7,8-tetrahydropterin produced by enzymic reduction (dihydrofolate reductase and NADPH) of 6-methyl-7,8-dihydropterin. *J. Chem. Soc., Chem. Commun.* **1980**, 334–336.
- (64) Peng, C.; Schlegel, H. B. Combining synchronous transit and quasi-Newton methods to find transition states. *Isr. J. Chem.* **1994**, *33*, 449–54.

- (65) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (66) Schwabe, T.; Grimme, S. Double-hybrid density functionals with long-range dispersion corrections: higher accuracy and extended applicability. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (67) Benach, J.; Atrian, S.; González-Duarte, R.; Ladenstein, R. The catalytic reaction and inhibition mechanism of Drosophila alcohol dehydrogenase: observation of an enzyme-bound NAD-ketone adduct at 1.4 Å resolution by X-ray crystallography. *J. Mol. Biol.* **1999**, *289*, 335–355.
- (68) Kavanagh, K. L.; Klimacek, M.; Nidetzky, B.; Wilson, D. K. Crystal Structure of Pseudomonas fluorescens Mannitol 2-Dehydrogenase Binary and Ternary Complexes. *J. Biol. Chem.* **2002**, *277*, 43433–43442.
- (69) Major, D. T.; Nam, K.; Gao, J. L. Transition state stabilization and alpha-amino carbon acidity in alanine racemase. *J. Am. Chem. Soc.* **2006**, *128*, 8114–8115.
- (70) Gready, J. E. Theoretical studies on the activation of the pterin cofactor in the catalytic mechanism of dihydrofolate reductase. *Biochemistry* **1985**, *24*, 4761–4766.
- (71) Rajagopalan, P.; Lutz, S.; Benkovic, S. Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates. *Biochemistry* **2002**, *41*, 12618–12628.
- (72) Sikorski, R. S.; Wang, L.; Markham, K. A.; Rajagopalan, P. T. R.; Benkovic, S. J.; Kohen, A. Tunneling and coupled motion in the E. coli dihydrofolate reductase catalysis. *J. Am. Chem. Soc.* **2004**, *126*, 4778–4779.
- (73) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (74) Brooks, B. R., III; C. L., B., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoseck, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (75) Sem, D. S.; Kasper, C. B. Enzyme-substrate binding interactions of NADPH-cytochrome P-450 oxidoreductase characterized with pH and alternate substrate/inhibitor studies. *Biochemistry* **1993**, *32*, 11539–11547.
- (76) Khavrutskii, I.; Price, D.; Lee, J.; Brooks, C., III. Conformational change of the methionine 20 loop of *Escherichia coli* dihydrofolate reductase modulates pKa of the bound dihydrofolate. *Protein Sci.* **2007**, *16*, 1087.
- (77) Stone, D.; Paterson, S. J.; Raper, J. H.; Phillips, A. W. The amino acid sequence of dihydrofolate reductase from the mouse lymphoma L1210. *J. Biol. Chem.* **1979**, *254*, 480–488.
- (78) Smith, D. R.; Calvo, J. M. Nucleotide sequence of the E. coli gene coding for dihydrofolate reductase. *Nucleic Acids Res.* **1980**, *8*, 2255–2274.
- (79) Murzina, N. V.; Gudkov, A. T. Invariant amino acid replacement affects the dihydrofolate reductase function and its gene expression. *Protein Eng.* **1990**, *3*, 709–712.
- (80) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U. K., 1987; pp 156–162.
- (81) MacKerell, A. D.; Bashford, D.; Bellott, Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (82) MacKerell, A. D.; Feig, M.; Brooks, C. L., III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
- (83) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (84) Gao, J. Toward a Molecular Orbital Derived Empirical Potential for Liquid Simulations. *J. Phys. Chem. B* **1997**, *101*, 657–663.
- (85) Sprik, M.; Klein, M. L.; Chandler, D. Staging - a sampling technique for the Monte-Carlo evaluation of path-integrals. *Phys. Rev. B: Condens. Matter* **1985**, *31*, 4234–4244.
- (86) Major, D. T.; Garcia-Viloca, M.; Gao, J. L. Path integral simulations of proton transfer reactions in aqueous solution using combined QM/MM potentials. *J. Chem. Theory Comput.* **2006**, *2*, 236–245.
- (87) Hwang, J. K.; Chu, Z. T.; Yadav, A.; Warshel, A. Simulations of quantum mechanical corrections for rate constants of hydride-transfer reactions in enzymes and solutions. *J. Phys. Chem.* **1991**, *95*, 8445–8448.
- (88) Major, D. T.; Gao, J. L. Implementation of the bisection sampling method in path integral simulations. *J. Mol. Graphics Modell.* **2005**, *24*, 121–127.
- (89) Cao, J.; Voth, G. A. A unified framework for quantum activated rate processes. I. General theory. *J. Chem. Phys.* **1996**, *105*, 6856–6870.
- (90) Gillan, M. J. Quantum simulation of hydrogen in metals. *Phys. Rev. Lett.* **1987**, *58*, 563.
- (91) Nam, K.; Gao, J.; York, D. M. An Efficient Linear-Scaling Ewald Method for Long-Range Electrostatic Interactions in Combined QM/MM Calculations. *J. Chem. Theory Comput.* **2005**, *1*, 2–13.
- (92) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.
- (93) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A: At. Mol. Opt. Phys.* **1985**, *31*, 1695.
- (94) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (95) Pu, J.; Gao, J.; Truhlar, D. G. Multidimensional Tunneling, Recrossing, and the Transmission Coefficient for Enzymatic Reactions. *Chem. Rev.* **2006**, *106*, 3140–3169.
- (96) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (97) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (98) Bieri, J. The crystal structure of 6-methyl-7,8-dihydropterine-mono-hydrochloride-mono-hydrate. *Helv. Chim. Acta* **1977**, *60*, 2303–2308.
- (99) Weber, R.; Viscontini, M. Über Pterinchemie 53. Mitteilung: Zur Konformation von 6-Methyl- und 5,6-Dimethyl-5,6,7,8-tetrahydropterin in wässriger Lösung. *Helv. Chim. Acta* **1975**, *58*, 1772–1780.
- (100) Poe, M.; Hoogsteen, K. 5,6,7,8-Tetrahydrofolic acid. Conformation of the tetrahydropyrazine ring. *J. Biol. Chem.* **1978**, *253*, 543–546.
- (101) Ganguly, A. N.; Bieri, J. H.; Viscontini, M. Über Pterinchemie. 77. Mitteilung: Das (6R,S)-5-Formyl-6-methyl-5,6,7,8-tetrahydropterin: Synthese, chemische und physikalisch-chemische Eigenschaften. *Helv. Chim. Acta* **1981**, *64*, 367–372.
- (102) Williams, T. C.; Strom, C. B. Tetrahydrobiopterin analogs: the solution conformations of 6-methyltetrahydropterin, 7-methyltetrahydropterin, and cis- and trans-6,7-dimethyltetrahydropterins as determined by proton nuclear magnetic resonance. *Biochemistry* **1985**, *24*, 458–466.
- (103) Yamada, T.; Liu, X.; Englert, U.; Yamane, H.; Dronskowski, R. Solid-State Structure of Free Base Guanidine Achieved at Last. *Chem.—Eur. J.* **2009**, *15*, 5651–5655.
- (104) Caminiti, R.; Pieretti, A.; Bencivenni, L.; Ramondo, F.; Sanna, N. Amidine N—C(N)—N Skeleton: Its Structure in Isolated and Hydrogen-Bonded Guanidines from ab Initio Calculations. *J. Phys. Chem.* **1996**, *100*, 10928–10935.
- (105) Zhong, H.; Bowen, J. P. Theoretical study of stereoselective reduction controlled by NADH analogs. *J. Mol. Graphics Modell.* **2005**, *24*, 1–9.

(106) Cummins, P. L.; Gready, J. E. Mechanistic aspects of biological redox reactions involving NADH 1: Ab initio quantum chemical structures of the 1-methyl-nicotinamide and 1-methyl-dihydronicotinamide coenzyme analogues. *THEOCHEM* **1989**, *183*, 161–174.

(107) Wu, Y. D.; Houk, K. N. Theoretical study of conformational features of NAD⁺ and NADH analogs: protonated nicotinamide and 1,4-dihydronicotinamide. *J. Org. Chem.* **1993**, *58*, 2043–2045.

(108) Almarsson, Ö.; Bruice, T. C. Evaluation of the factors influencing reactivity and stereospecificity in NAD(P)H dependent dehydrogenase enzymes. *J. Am. Chem. Soc.* **2002**, *115*, 2125–2138.

(109) Karle, I. The crystal structure of N-benzyl-1,4-dihydronicotinamide. *Acta Crystallogr.* **1961**, *14*, 497–502.

(110) Glasfeld, A.; Zbinden, P.; Dobler, M.; Benner, S. A.; Dunitz, J. D. Crystal structures of two simple N-substituted dihydronicotinamides: possible implications for stereoelectronic arguments in enzymology. *J. Am. Chem. Soc.* **1988**, *110*, 5152–5157.

■ NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on September 14, 2011. A change has been made to the caption of Figure 4. The correct version was published on September 21, 2011.

Gating and Intermolecular Interactions in Ligand-Protein Association: Coarse-Grained Modeling of HIV-1 Protease

Myungshim Kang,[†] Christopher Roberts,[†] Yuhui Cheng,[‡] and Chia-en A. Chang^{†,*}

[†]Department of Chemistry, University of California, Riverside, California, United States

[‡]Pacific Northwest National Laboratory, Richland, Washington, United States

 Supporting Information

ABSTRACT: Most biological processes are initiated or mediated by the association of ligands and proteins. This work studies multistep, ligand–protein association processes by Brownian dynamics simulations with coarse-grained models for HIV-1 protease (HIVp) and its neutral ligands. We report the average association times when the ligand concentration is 100 μM . The influence of crowding on the simulated binding time was also studied. HIVp has flexible loops that serve as a gate during the ligand binding processes. It is believed that the flaps are partially closed most of the time in its free state. To accelerate our simulations, we fixed a part of the HIVp and reparameterized our coarse-grained model, using atomistic molecular dynamics simulations, to reproduce the “gating” motions of HIVp. HIVp–ligand interactions changed the gating behavior of HIVp and helped ligands diffuse on HIVp surface to accelerate binding. The structural adjustment of the ligand toward its final stable state was the limiting step in the binding processes, which is highly system dependent. The intermolecular attraction between the ligands and crowder proteins contributes the most to the crowding effects. The results highlight broader implications in recognition pathways under more complex environment that considers molecular dynamics and conformational changes. This work brings insights into ligand–protein associations and is helpful in the design of targeted ligands.

INTRODUCTION

In biological systems, many processes such as the immune response, signal transduction, and metabolism are initiated or mediated by ligand–protein associations.^{1–3} Several factors affecting ligand–protein associations include conformational changes of a molecule, intermolecular interactions, gating effects, molecular crowding, and solvent effects.⁴ Notably, proper biomolecular function depends on a balance between the timing and duration of the ligand–protein interactions. The association rate constant is one of the most important properties that describe the movement of molecules in solvent. The diffusion-limited rate is the upper limit of the ligand–protein association rate,⁵ and many ligand–protein systems exhibit association rates slower than the diffusion-limited association rate. Understanding ligand–protein associations provides insight into regulating protein functions and can lead to practical applications such as drug design.

If a ligand and a protein are each approximated as spheres, and their surfaces are perfectly reactive, then the diffusion-limited rate constant is given by the classical Smoluchowski equation $k_{\text{on}} = 4\pi RD$, where R is the sum of the radii and D is the relative translational diffusion constant.⁶ The expression thus yields rates of 10^9 – 10^{10} $\text{M}^{-1}\text{s}^{-1}$ for most ligand–protein associations. However, proteins commonly have a highly anisotropic binding site, and molecules are not simple spheres, resulting in slower associations in a complex cellular environment. Although the ligand–protein association can be complicated, it may be simplified into 3 steps:^{7–9} the initiation step, characterized by the free diffusion of unbound ligand and protein molecules; a second step involving several intermediates to rearrange ligand, protein, and water molecules when the two molecules meet; and

the third step finalizing the binding process and resulting in the bound complex.

The free molecules and bound complex may be captured by high-resolution experimental structures, but in most cases, the second step in the process is poorly understood. During the second step, the ligand and protein optimize their interactions by changing their conformational preference or inducing new conformations.^{10–12} Protein structural fluctuations also affect binding, and the binding sites of proteins can be occluded by motions of protein loops. Such flexible loops or flaps may be described as a “gate” that can open and close to influence ligand binding. An early analysis from McCammon and Northrup revealed two limiting cases: fast and slow gating.^{13–15} In fast gating, the gate opens and closes much faster than the rate of escape of the ligand from the protein proximity into the solvent. The protein can thus be viewed as “always-open”. In slow gating, the gate opens and closes much slower than the rate of escape of the ligand. Therefore, the successful binding rate is that of the always-open protein multiplied by the opening fraction of the gate. As compared with the second step, the initiation step is relatively well studied in the dilute environment, so theories and analytical models are available for diffusion-controlled ligand–protein associations.⁴ However, cells are not dilute environments, and approximately 20%–40% of the cytoplasmic volume is occupied by biomolecules or small compounds.¹⁶ Simple analytical models may not describe the large macromolecular crowding effects.

Received: July 13, 2011

Published: August 26, 2011

To study the molecular encounter process, we employed the model system of human immunodeficiency virus type 1 protease (HIVp). HIVp has been one of the major targets for AIDS treatments because it is essential in the viral replication cycle. The enzyme cleaves the viral polyproteins at the active site to make the replicated virus mature and become infectious.^{17,18} HIVp has flexible flaps over the binding site that open and close to “gate” ligand binding.¹⁹ Some HIVp inhibitors have similar binding affinities, i.e., the dissociation constant K_d but quite different kinetic behaviors in the association and dissociation rate constants k_{on} and k_{off} .^{20,21} Some mutations may be more sensitive to the kinetics, and elucidation of the HIVp–ligand association may aid in the drug discovery process.

In this work, we simulated ligand–protein association times using a coarse-grained (CG) model and the Brownian dynamics (BD) algorithm to connect the initial and final steps of binding processes. Each residue of HIVp is represented by one bead and the ligands are modeled by 5 or 6 beads. To accelerate our simulations, we fixed a part of HIVp and reparameterized our CG model by atomistic molecular dynamics simulations to resemble the motions of a fully flexible HIVp. To study the environmental crowding effects, we coarse-grained glutathione S-transferase P (GSTP1) as our crowder protein and as in HIVp–ligand interactions, our ligand had both attractive and repulsive interactions with the crowders. We found that the ligand–protein interactions alter the flap dynamics from slow gating to fast one. The structural adjustment of the ligand toward its final bound state was the limiting step in the binding process, which was highly system dependent. The intermolecular attractions between the ligand and crowders contributed the most to the crowder effects, as well.

METHODS

Protein and Ligand Structure. The HIVp structural coordinates were obtained from the protein data bank (PDB code 1HHP).²² One of the most abundant proteins in the human cell, GSTP1, (PDB code 3DGQ) was selected as the crowding protein.²³ The structures of the ligands XK263, ritonavir, and saquinavir were also obtained from the protein databank (PDB codes 1HVR, 2B60, and 3EKQ, respectively).^{24–26}

Coarse-Grained Model of Proteins. Various CG approaches have been developed to overcome the limitation of atomistic molecular dynamics simulations in terms of time scales as well as the system sizes.^{27,28} Here, we used CG models to represent HIVp, the crowding protein, and the ligands. In the CG model, a single interaction center, placed on the C_α position, was used to represent each amino acid.^{29–31} For the five charged amino acids, appropriate formal charges were assigned. The detailed procedure for the coarse graining from a collection of pdb files and the all-atom representations to the one-bead representation was previously described.³²

The coarse-grained potential energy function is defined as follows:

$$U = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihe}} + U_{\text{elec}} + U_{\text{vdw}}^{\text{intra}} + U_{\text{vdw}}^{\text{inter}} \quad (1)$$

When one of a pair of beads was fixed, the force between the pair was applied only on the flexible bead to avoid a possible collision. Although no detailed solvent model was used, a distance-dependent dielectric constant ($\epsilon = 4r_{ij}$) was used for compensation to avoid unrealistic Coulombic interactions.

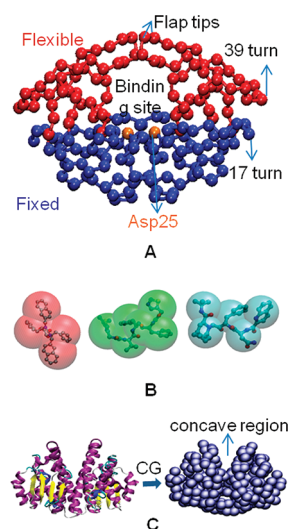


Figure 1. Coarse-grained models of HIV protease (HIVp), ligands, and GSTP1. (A) Model of HIVp. The flexible part and fixed part are in red and blue, respectively. The binding site has 2 aspartic acids (orange). (B) Models of ligands. XK263 (red, left), saquinavir (green, middle), and ritonavir (blue, right). (C) Cartoon representation (left) and model of glutathione S-transferase P (GSTP1) (right). GSTP1 has a concave region on the surface.

CG models for the ligands were also generated. The ligands were manually parametrized with interaction centers on the various functional groups in each structure (Figure 1). The net charge of each ligand is zero, but partial charges were applied to each bead (for details, see the Supporting Information).

For further acceleration of the simulation, the 108 beads distant from the highly mobile flaps of HIVp were fixed in their position. However, after fixation, the flaps did not open as widely as the free protein, which suggests possible correlations between the fixed and flexible parts. In fact, previous BD simulations suggested a correlation between the 17 and 39 turns of HIVp.³³ To properly reproduce the intramolecular motions of the flaps, we introduced modifications in angles and dihedrals. From previous atomistic molecular dynamics (MD) simulations,³⁴ the angles and the dihedrals in the open conformations were compared with those in the closed as well as transition conformations. The angles and dihedrals with significant differences were selected and their parameters were modified to fit the distributions from the atomistic MD simulation, based on the Boltzmann inversion^{31–33,35} (see Supporting Information for further details).

Brownian Dynamics. Simulations involved the use of the modified UHBD software package as previously described.^{36,37} The maximum simulation time for each run, without crowding, was 30 μs because more than 99.5% of runs bind within this time. With crowding, the maximum simulation time was set to 100 μs for a similar success rate of binding. The time step was 50 fs for all runs and the trajectory was saved every 1 ns without crowding and 10 ns with crowding. The viscosity of water η was set to 1 cP, which corresponds to a water temperature of 293 K. The total system is spherical in nature, with a radius of 160 Å. This represents the volume necessary to obtain a single ligand concentration of 100 μM , which corresponds to experimental concentrations used for comparison.

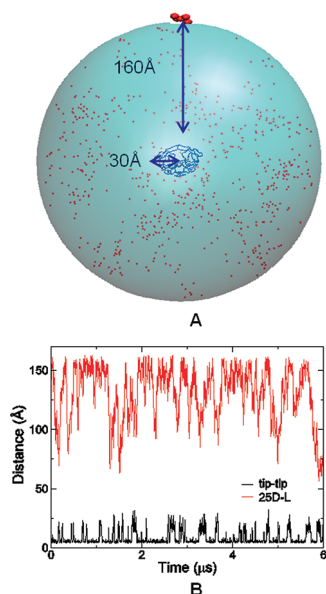


Figure 2. Diffusion association model. (A) HIVp, with a radius of about 30 Å, is centered in the spherical system of radius 160 Å. A ligand starts to diffuse in a random position on the surface of the sphere. Red dots represent the trajectory of an example ligand. (B) The tip to tip distances and distances between a ligand and the active site by simulation time.

Each simulation began with a single ligand placed in a random position on the surface of the bounding sphere, 160 Å from the point of origin to a central bead of the ligand (Figure 2). The ligand follows a Brownian trajectory over the time step given. If the trajectory were such that the ligand escaped the bounding sphere, a periodic boundary condition-like conversion was used to place the ligand at a diagonally opposite direction of the 160 Å sphere, with an additional pull of 2% total distance toward the origin to ensure that the ligand was inside the sphere. This condition ensures that the ligand concentration remains constant. Additionally, a binding event occurred if 2 conditions were satisfied: the ligand diffuses within 11 Å of the active site, and the protease conformation is in the closed position. These criteria were defined by the distance between the ligand and Asp25 and the distance between the tips of each flap of the protease, respectively.

Calculation of Translational Diffusion Coefficients. The Einstein formula was used to calculate the translational diffusion coefficients for the ligand molecules:

$$D_{trans} = \frac{\langle \delta r^2 \rangle}{6\delta t} \quad (2)$$

where δr is the distance traveled in 3 dimension space during the time interval δt . The equation describes the ensemble average of the center of mass Cartesian trajectory over the time interval δt . The time interval selected for the ligand diffusions were 1 and 10 ns. The longer time step represents the time needed to average out anomalous diffusion, and the shorter time step was used to ensure statistical accuracy of the method.

Crowding Model. To assess the degree to which crowding in a cytoplasm can affect the ligand–protein association, we built a crowding model. We selected GTP1 as a crowder and built a CG model for it as we did for HIVp. We fixed the crowder protein models during the simulation. To reduce the computational penalty of additional structures in the system, we removed all

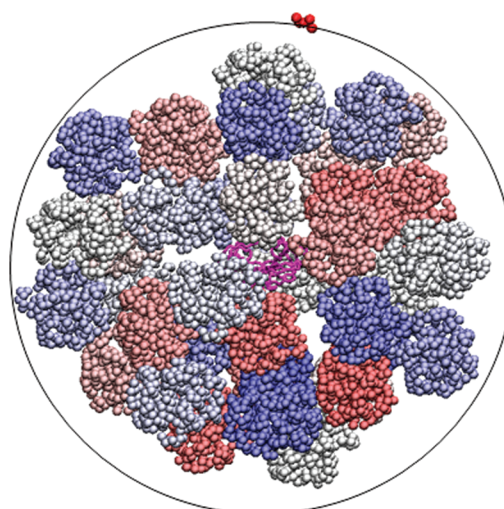


Figure 3. 40% crowding model. HIVp (magenta) is centered in the sphere and each run begins with a ligand (red) on the surface of the sphere. The crowder proteins were inserted in our sphere randomly, but we reserved a radius around HIVp to ensure that the flaps had full range of motion.

nonsurface beads, because they would have negligible non-bonded interactions with the ligand. Full attractive and repulsive van der Waals and electrostatic interactions were enabled. We inserted the crowder proteins in our sphere randomly but reserved a radius around HIVp to ensure that the flaps had full range of motion. Two crowding models were simulated, with crowder protein concentrations of 20% and 40% of total volume (see Figures 3 and S8). These concentrations represent the cytoplasmic conditions.

RESULTS

Validation of the Coarse-grained Brownian Dynamics Model. We analyzed the trajectories from CGBD simulations of HIVp–ligand association processes to study gated diffusion-controlled encounters. The protein's two polypeptide “flaps” that cover the active site at the dimer interface were in the open position for most of the binding process. To keep the correct flap open/closed fraction and simulation time scale, the HIVp was flexible even when the ligand was far from the protein (Figure 2). Because the simulations are computationally intensive (2 h/1 μs using 1 CPU), we fixed the atoms in the bottom half of HIVp structure. However, fixing half of the protein diminished the possible correlation between the 17 and 39 turns (Figure 1),³³ resulting in a more rigid protein with a small open fraction, as compared with the same CG model with a fully flexible protein. Therefore, we restored the protein's internal motion by modifying a few parameters of angles and dihedral angles based on the open conformations, as detailed in the Methods section. Our partially fixed wild-type HIVp has about 15% open fraction, with mean open and closed times of ~30 ns and ~150 ns, respectively (see Supporting Information for details).

In our BD simulation, the Ermak-McCammon algorithm did not include the hydrodynamic tensor for N particles/beads in order to speed up the calculations, as computing the tensor results in a run time of $O(N^3)$.³⁸ Because the $3N \times 3N$ diffusion tensor determines the correlated random displacements for N particles/beads, we examined the effect of neglecting this

Table 1. Average Association Times (μs) of the Ligands XK263, Ritonavir and Saquinavir to HIV Protease (HIVp) in Simplified Coarse-Grained (CG) Models^a

HIVp model	binding criteria	XK263	ritonavir	saquinavir
(A) no interaction	surface of 30-Å sphere	4.49 (1000)	4.67 (1000) 4.73 (1600)	3.97 (1000)
(B) fixed closed CG	surface of 30-Å sphere	3.18 (1000)	3.83 (1000)	3.32 (1000)
(C) fixed open CG	$D_{\text{protein-ligand}} \leq 11 \text{ \AA}$	5.37 (131)	6.31 (131)	5.04 (131)
(D) Partially flexible CG	$D_{\text{protein-ligand}} \leq 11 \text{ \AA}$, $D_{\text{tip}} \leq 7 \text{ \AA}$	4.70 (2140) 4.60 (6879)	5.49 (2140)	4.63 (2140)
exp. k_{on}^b ($\text{M}^{-1} \text{ s}^{-1}$)		$\sim 10^9$	$\sim 10^6$	$\sim 10^5$
exp. ΔG^c (kcal/mol)		-13.5	-9.3 ~ -14.4	-8.5 ~ -13.0

^aThe numbers of runs are in the brackets. The standard deviations are in Table S2 of the SI. ^bRef 55. ^cRef 21.

hydrodynamic interaction when diffusing our ligands. When we freely diffused the ligands using only one bead with the approximated radius as in eq 2, the translational diffusion coefficients from our BD algorithm of XK263, ritonavir and saquinavir were 2.2×10^{-6} , 2.1×10^{-6} , and $2.2 \times 10^{-6} \text{ cm}^2/\text{s}$, respectively, in the dilute solution. These coefficients are in good agreement with the analytical values, thus our algorithm accurately diffuses a sphere solute in free solution. When using the multibead models (see Figure 1), the diffusion coefficients are 0.9×10^{-6} , 0.8×10^{-6} , and $0.9 \times 10^{-6} \text{ cm}^2/\text{s}$, respectively. In addition, using 1 or 10 ns as the observed time interval (δt in eq 2) did not affect the computed diffusion constants. Although the simulated diffusion coefficients are 2 to 3 times slower than the analytical values because of the more complicated molecular shape and simplified hydrodynamic interactions, the coefficients keep the same order as those computed from the simple sphere models. Neglecting the hydrodynamic interactions may reduce the diffusion coefficients; however, because our ligands have only 5 to 6 beads, the effects are not significant.

Association Time for Ligand Binding to Simplified HIVp.

On the basis of experimental assays, our ligand concentration was set to $100 \mu\text{M}$, which is equivalent to the diffusion of one free ligand within a sphere of 160 \AA while HIVp is located in the center of the sphere. Instead of computing the association rate constant with a probability-based algorithm such as Northrup-Allison-McCammon (NAM),³⁹ we computed the association time, the time it takes each ligand to reach the binding site of HIVp. Each run began with a ligand on the surface of the 160-\AA sphere, and the simulation stopped when the binding criteria was satisfied. We studied factors such as the size of the active site, intermolecular interactions between the protein and ligands, and intramolecular “gating” motion of the protein, so several simplified models were also tested (Table 1 and Figure 4). We started with the simplest of these models, in which HIVp is approximated by a 30-\AA radius sphere without any interactions with the ligands, and the binding is considered successful when a ligand reaches anywhere on its surface. As shown in Table 1, all of the ligands needed nearly $4 \mu\text{s}$, on average, to associate with the sphere protein, and the standard deviation was as large as the association time (see Supporting Information for details). With eq 2 and a diffusion coefficient of $\sim 2 \times 10^{-6} \text{ cm}^2/\text{s}$, the time required for diffusing 160 \AA in one dimension takes only $\sim 0.6 \mu\text{s}$, with no analytical association time available. Therefore, we used the association time obtained from this simplest model as our reference.

To study how the intermolecular interactions may affect the association time, we replaced the 30-\AA sphere by HIVp with a

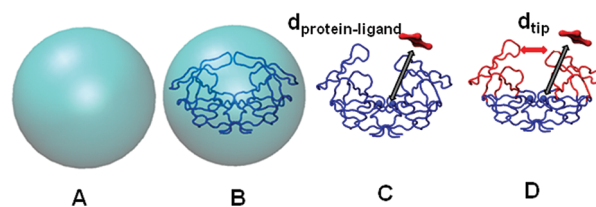


Figure 4. Simplified models for HIVp. (A) 30 \AA radius sphere. Ligands “bind” at the surface of the sphere. There is no intermolecular interaction between the protein and the ligand. (B) Fixed HIVp CG model with closed conformation in a 30-\AA radius sphere. The intermolecular interaction is introduced, but ligands still “bind” at the surface of the 30-\AA radius sphere. (C) Fixed HIVp CG model with open conformation. The ligands bind when they are within 11 \AA of the Asp25 of the active site. (D) Partially flexible HIVp CG model with both fixed (blue) and flexible parts (red). The ligands bind when they are within 11 \AA of the Asp25 of the active site and the flaps are closed. (also see Table 1).

closed conformation and again considered the binding successful when the ligand reached 30 \AA from the center of the protein. Because our ligands are neutral, without long-range electrostatic steering, van der Waals attractions slightly decreased the association time by $\sim 17\%$ to 20% . We further examined the geometric effects by obtaining the association time for ligands binding to the active site of HIVp with an open conformation. Because the protein is fixed, once the ligand is within 11 \AA of the active site, it is considered bound. This simulation also provides insight into how the ligands would bind to HIVp if there were no “gating” motion. The geometry constraint leads to an increase in mean binding times by 1.5 to 1.7-fold, although the binding site occupies only about 10% of the protein surface. Purely probabilistic models account for such steric constraints by multiplying the Smoluchowski rate for uniform spheres by the probability that, in a random encounter, the two molecules are properly bound. Although the binding site is decreased 90%, the mean binding times did not increase as much, which suggests that the association is accelerated by surface diffusion. The initial contact of the ligand can be anywhere on the surface of HIVp. Instead of immediately leaving HIVp, the ligands diffused laterally along the HIVp surface to search for the correct orientation and protein binding site. The diffusion constants were substantially reduced (see Table 2), but the lateral diffusion largely decreases the association time because the ligand does not need several hundred attempts to reach the binding site of HIVp. Interestingly, diffusion along the z -axis, corresponding to the longest horizontal dimension across the HIVp, was faster for all ligands.

Table 2. Translational Diffusion Coefficients, D_T ($\times 10^{-6}$ cm²/s), 1000 Samples, for XK263, Ritonavir and Saquinavir with HIVp

model		XK263	ritonavir	saquinavir	δt (ns)
diffusion through space					
1 bead	approximate radius (Å)	7.3	8.0	7.5	
	D_T	2.25	2.16	2.29	10
		2.24	2.06	2.18	1
multibead	D_T	0.93	0.75	0.93	10
		0.90	0.76	0.90	1
lateral diffusion on the surface of the protein					
multibead	D_T	0.34	0.26	0.34	1
		0.08	0.07	0.09	10
	D_X^a	0.45	0.09	0.19	1
		0.11	0.02	0.06	10
	D_Y^a	0.13	0.34	0.39	1
		0.02	0.11	0.07	10
	D_Z^a	0.42	0.34	0.44	1
	0.11	0.08	0.13	10	

^a D_X , D_Y , and D_Z represent 1-D translational diffusion coefficients in x, y, and z axes, respectively. See also Figure S4.

Of note, use of a different observation time interval (δt in eq 2) results in different translation coefficients when the ligand diffuses along the protein surface.

Association Pathways and Motions of HIVp. As summarized in Table 1, our CG models for HIVp and ligands did not show large differences between the ligand–protein association times. During the simulations, the flaps of HIVp could open and close, and the final bound state was defined by two distances: the distance between the tip of the flaps and that between the center bead of a ligand and Asp25 of the protein. After the gating motions of the flaps were introduced, the average bindings were approximately 14% faster than those with a fixed open CG model, while the gate only opened $\sim 15\%$ in the free protein. For example, as shown in Figure 2, before the ligand approached the protein, the flaps opened many times, with the open fraction $\sim 15\%$. According to previous studies, these flap motions may be considered “slow gating”; thus, the rate for successful binding should be reduced significantly because only 15% of the attempted bindings can be successful. Therefore, the association time was expected to increase considerably, since the ligand may leave the protein easily when the flaps are not open. However, the ligand can wander around the enzyme until it reaches the neighborhood of the binding site (see Figure S4), and the ligand–protein interactions modify the flap dynamics, making the opening more frequent to facilitate ligand binding.

Because the initial positions of the ligands are far from HIVp, the ligand’s initial position and orientation did not affect the general binding pathway. Figure 2 shows the density of the ligands from simulations with different starting positions on the 160-Å sphere, which confirms that all ligands explore the surface of the enzyme. Although studies have suggested that the surface of the “hinge” region of the flaps of HIVp may be a potential ligand binding site,^{40,41} all three ligands did not show higher density near the “hinge”, but tended to stay longer near

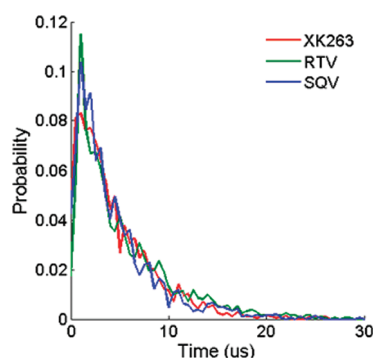


Figure 5. Histogram for the association time for each ligand to bind to HIVp.

the bottom of HIVp, where two cavities can be observed. To determine whether the trapping is an artifact from the fixation of the bottom portion of the enzyme, we also carried out 25 BD simulations with a fully flexible HIVp. In these simulations, ligands also stayed longer in the bottom part of the protein (see Figure S4 of the SI for the density plot), thus the rigid part of HIVp does not produce artificial intermolecular attractions. Although our ligands formed a very stable ligand–protein complex in our coarse-grained model, occasionally the ligands left the binding sites and returned again, as seen in the atomistic simulation.

With our binding criteria, the association time for XK263 was the shortest, but the times were similar for all three ligands (see Table 1). Note that although instruments for measuring association or dissociation rate constants, such as BIAcore, have a data collection rate of about 1–100 Hz, we set a maximum of 30 μ s for our simulations. Figure 5 illustrates the probability of binding at different cutoff times, and XK263 shows slightly higher probability of binding. For example, within 0.6 μ s, the analytical time for diffusing a 7.3-Å spherical ligand over 160 Å, the probability of binding for XK263 is only about 20% larger than that for ritonavir and saquinavir. In contrast, the experimental association rate constant for XK263 is about 1000 times larger than those for ritonavir and saquinavir. However, although HIVp–ligand interaction energies and association time are similar for all ligands (see Figure 6), XK263 has much smaller fluctuations in the active site of the protein. Ritonavir and saquinavir did not leave the binding site after they reached the bound state, but they tumbled more frequently and the flaps opened more often. In a few cases, ritonavir and saquinavir show smaller ligand fluctuations and very few flap openings in the bound state, so some may reach the final bound state, and our CG model can form the final bound state. The movement of both ritonavir and saquinavir suggest that they need significantly longer time than XK263 to adjust the conformation to find the final bound state.

Our CG model uses simple criteria to determine successful binding; however, the final bound conformation may need atomistic details, including waters, to fully optimize the final bound complex. Of note, XK263 does not need water molecules, but other peptidomimic ligands such as ritonavir and saquinavir have one crystal water in the binding site in their experimental structures. This work focused on only a wild-type protein. Mutants that change the flap motions of HIVp may affect the ligand–protein association, which is not discussed here.

Ligand Binding in Crowded Environment. To study ligand–protein encounters in the crowded environment of cells, we selected

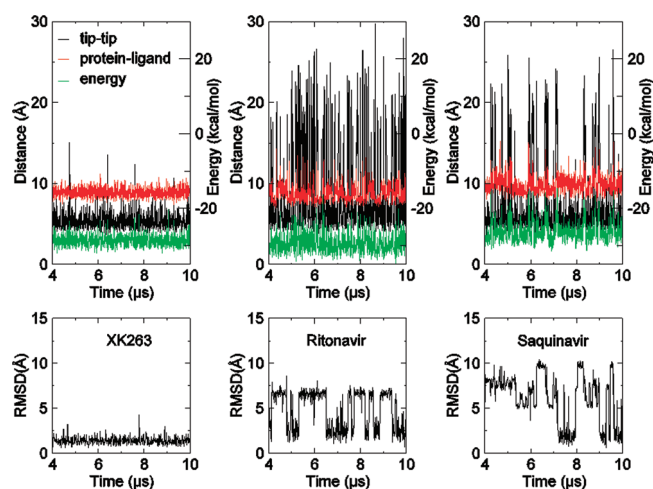


Figure 6. The different behaviors of the ligands at the binding site. The distance between the protein (Asp25) and the center of each ligand, the tip to tip distance, and the ligand–protein interaction energy (top), and the root-mean-square deviation (rmsd) of each ligand after they “bind” (bottom).

Table 3. Mean Association Times, t_{on} , and Diffusion Coefficients, D_T , of XK263 in Crowding Models

interaction	repulsion + attraction	repulsion only		
concentration (%)	20	40	20	40
$t_{\text{on}, c\%}$ (μs)	11.0	19.0	4.8	4.9
$t_{\text{on}, c\%}/t_{\text{on}, 0\%}^a$	2.3	4.0	1.0	1.1
$D_T(10^{-6}\text{cm}^2/\text{s})^b$	0.41	0.24	0.74	0.75
no. of samples	300	300	300	300

^aThe ratio of the mean association time with $c\%$ crowding, $t_{\text{on}, c\%}$, to that without crowding, $t_{\text{on}, 0\%}$. ^bThe dt for the translational diffusion coefficient is 10 ns.

and coarse-grained one of the most abundant proteins in human cells, GSTP1, as a crowder agent. All GSTP1 molecules were held rigid and had intermolecular attractive and repulsive interactions with our ligands (eqs S6 and S7 of the SI). Typically, 20% to 40% of the cytoplasmic volume is occupied by macromolecules, and we had two crowding systems: GSTP1 in 20% (20 proteins) and 40% (38 proteins) of excluded volume concentration in our 160-Å sphere, where GSTP1 were randomly placed in the sphere (see Figures 3 and S8 of the SI). GSTP1 is approximately 2.2- times larger than HIVp, and a few small shallow cavities can be observed on the protein surface (see Figure 1C). Previous work showed that crowded conditions can affect the internal dynamics of HIVp.⁴² Therefore, we placed GSTP1 far enough, with the closest distance between HIVp and GSTP1 ~ 26 Å, to avoid any influence of crowding on HIVp internal dynamics, and the intermolecular interactions between HIVp and GSTP1 were also turned off. We also ran BD simulations for HIVp with the crowders to ensure that the fractions of open/closed flaps remain unchanged. Since we fixed GSTP1 proteins during our simulations, two different crowder configurations with various protein orientations were used, and the ligand binding time did not show noticeable difference from the simulations (data not shown).

Table 3 shows the average binding time of different ligands for both crowding environments. The binding is slowed by two- to 4-fold with crowding, which is similar to other simulations.

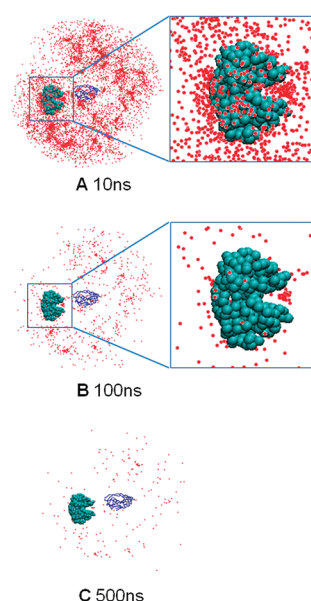


Figure 7. The density of XK263 with 20% crowding. Snapshots at 10 (A), 100 (B), and 500 ns (C). One GSTP1 is shown in the extended illustration (right) to show increased density near the concave region on the surface of GSTP1.

For example, the association time for XK263–HIVp binding increased from 4.7 to 19.0 μs with 40% of the volume occupied by GSTP1. No simple relation was found between binding time and crowder concentration—the attraction between the ligand and GSTP1 slowed the HIVp–ligand association time, but the reduction in ligand diffusion space could also accelerate HIVp–ligand binding. Figure 7 illustrates the density of XK263 when encountering the protease from various starting positions on the 160 Å sphere. The snapshots saved every 10, 100, and 500 ns show an increased density near the concave regions of the GSTP1 surface. With snapshots saved every 500 ns, the clusters of ligand distribution in those regions are less noticeable, thus the drug may be temporarily trapped by the crowders for up to a few hundred ns. To test the effects of intermolecular attractions to HIVp–ligand encounters, we turned off the intermolecular attractive forces between the ligand and GSTP1 and kept only the repulsive potential. The association time was similar to that in dilute solution.

Crowders occupying space may effectively decrease the volume available for the ligand when it diffuses toward the enzyme, which can also be considered as increasing the ligand concentration. We therefore reduced the size of the sphere to 149 and 135 Å to make the ligand concentrations equivalent to the effective concentrations when the space is 20% and 40% excluded by GSTP1, respectively. The association time is reduced 33% and 50%, which showed a positive correlation between the binding time and the space reduced.

DISCUSSION

General Comments. This work aimed to provide a more complete description of ligand–protein association behavior with the occurrence of significant protein conformational changes during the ligand–protein encounter. We also studied the influence of crowder proteins on the simulated binding time, which provides insights into molecular binding in the crowded

environment. Instead of computing the association rate constants, we reported the mean association times (conceptually similar to inverse rates) for ligands in a dilute solvent and crowded environment. The ligand concentration was set to 100 μM . As previously stated, a 1 M ligand concentration is equivalent to 1 molecule/1660 \AA^3 , so our model freely diffuses one ligand in a volume inside a 160 \AA sphere. If a ligand diffuses outside the sphere, then we ensured a consistent concentration by placing a ligand in the opposite direction of the 160 \AA sphere, as mentioned in the Methods. One can also place multiple ligands in a larger volume for a 100 μM ligand concentration to simulate the effects of ligand–ligand interactions on association times. Because our ligands are neutral, without noticeable long-range electrostatic interactions, we assume that the influence of these interactions can be ignored. However, for charged ligands, or in substantially high ligand concentration, the ligands may interact with each other or compete for the binding site of the protein.⁴³

Unlike a natural substrate that will be cleaved and then the product released, the chemical compound stays in the binding site. If the HIVp concentration is not very low and the dissociation rate is slow (e.g., $< 10^{-2} \text{ s}^{-1}$), then a ligand may already bind in the binding site of HIVp when our ligand diffuses close to the binding site. This study assumes that the HIVp concentration is significantly lower than the ligand concentration, so when the ligand diffuses near HIVp, the enzyme is in a free state and no compound occupies the binding site. The simulation times are sufficiently long for the ligands to have enough time to reach the binding site. More than 99.5% of ligands could bind to the protein within 30 μs in the dilute system. However, to reach the same success rate, we needed a 100- μs simulation time when crowders are present.

At the given ligand concentration, the association rate constant can conceptually be converted to the approximate association time and vice versa. For example, the association time of XK263 estimated from the experimental association rate constant at the ligand concentration of 100 μM is $\sim 10 \mu\text{s}$. The association rate constant computed by the Smoluchowski expression using the diffusion coefficients from a single sphere, $D_T = 2.25 \times 10^{-6} \text{ cm}^2/\text{s}$ for XK263, yields $k_{\text{on}} \approx 6 \times 10^9 \text{ M}^{-1} \text{ s}^{-1}$. The approximate association time based on this k_{on} is $\sim 1.7 \mu\text{s}$. Notably, the diffusion-control association rate constant for two molecules without intermolecular interactions is $10^9 - 10^{10} \text{ M}^{-1} \text{ s}^{-1}$ (see Smoluchowski equation in the introduction). Our calculations of XK263 suggest that the gating effects do not remarkably slow down the association time, and the association is close to diffusion-control processes. Ligands may diffuse away when they reach the nonbonding site regions of HIVp or when the flaps are closed. Nevertheless, in most cases, XK263 can diffuse on the HIVp surface to wait for the flaps to open or bind from the side when the flaps are not fully open. In addition, the ligand–protein interactions can facilitate the opening of the flaps, which alters the flap dynamics from “slow gating” to “fast gating” for HIVp. Previous modeling showed that the flap dynamics leads to “slow gating”.¹⁹ According to the gating theory, a slow gating effect suggests that the gated association rate constant is simply the ungated rate constant times the probability of the gate opening.^{14,15} Although the opening of the flaps is not large, the space is spacious enough for small drug binding. The flaps can open fast enough to not affect the binding of the ligand (fast gating). Note that although our ligands are small molecules, natural substrates are large polypeptides that may need fully open flaps. Consequently,

the slow gating feature may remain depending on the ligand. The modifications of flap motions due to intermolecular interactions are consistent with previous studies.^{33,34,44} In addition, the facilitation of binding by diffusion of ligands along the protein surface has been previously suggested.⁴⁵

As expected, the analytical results from eq 2 are similar for all ligands tested because they have similar sizes and diffusion coefficients. Surprisingly, the mean association times from our simulations were also similar for all ligands, which is inconsistent with experimental data. While the three ligands have similar binding affinities, the association rate constants are a few orders slower in saquinavir and ritonavir (Table 1). Before the ligands reach the binding site, their association processes were similar, and they can all induce flap opening. All ligands may become temporarily trapped in the bottom of the HIVp structure, with subsequent diffusion along the protein surface. The attraction energies for nonbonded van der Waals interactions (eq S6 of the SI) may be overestimated, so once a ligand forms an attractive potential with HIVp, it stays with the protein, thus resulting in a rate constant near that of diffusion-limit association. However, our ligands do not stay with the crowders, which use the same nonbonded potential energy function as those used in HIVp. Therefore, overstabilizing ligand–protein interactions is not likely. During our simulations, sometimes a ligand left the enzyme after encountering the surface but returned to attempt binding two or three times. As a result, the standard deviations of the association times are all large (see Figure S5 of the SI).

After the ligand diffused into the binding site and the flaps of HIVp were fully closed, XK263 showed less fluctuation in the protein pocket, but saquinavir and ritonavir kept tumbling in the binding site. We also carried out several 200 to 300 μs long simulations, and saquinavir and ritonavir could reduce their fluctuation after $>100 \mu\text{s}$ in a few runs. Therefore, if we include the ligand stability in the HIVp binding site as a criterion to define the bound state, then saquinavir and ritonavir require relatively long association times as compared with XK263. Despite the similar intermolecular interactions between HIVp and different ligands in our CG model, a few details crucial to stabilize the ligand-bound state cannot be fully captured by the model. As previously mentioned, one detail is the role of water molecules in the binding-site to stabilize the ligands. During the binding processes, the association time spent to replace and recruit the bound water can differ greatly between cyclic urea ligands (e.g., XK263) and peptidomimetic drugs (e.g., saquinavir and ritonavir). As a result, to thoroughly study the second step of the ligand binding mechanism, an atomistic force field with an explicit water model may be necessary. Nevertheless, our CG model with BD simulations is an efficient method to gain insight into and an overview of the binding processes in both dilute and crowded environments.

Binding in Crowded Environments. Under crowding conditions, as in cytoplasm, the effective molar concentration of a ligand is increased with decreased volume of the solution. The macromolecules may interact nonspecifically with the ligand and the target protein in vivo. Previous publications showed that crowders can affect protein dynamics and folding.^{11,42,46–51} Even though the hydrodynamic interactions have been suggested to be significant in modeling crowding and macromolecular motion,^{52,53} the hydrodynamic interaction is insignificant because the ligands are small and crowders are held fixed. This work focused on the crowding effects contributed by the excluded diffusion volume and crowder–ligand interactions. We first increased the ligand

concentration by reducing the diffusion volume to a space in the 145 and 139 Å sphere, and the association time decreased with higher ligand concentrations. Without intermolecular attraction forces, we expected that the association time may also be shorter because the available space for ligand diffusion was cut by 20% to 40% (volume occupied by the crowders).⁵⁴ However, the association time is the same with and without crowders, although the diffusion space for XK263 was reduced by 20% and 40%. The volume was reduced because of the volume excluded by the crowders, and the translational diffusion coefficients were also reduced, so the crowding actually slows down the diffusion, but it is canceled out by the excluded volume effect. Thus, when only repulsive forces are considered, excluded volume effects are not significant for diffusing small molecules to a protein target. The distance that the ligand needs to travel may be the main determinant. In contrast, if we consider all intermolecular attractions, then the association times increase 2.3-fold with 20% crowders and nearly 4-fold with 40% crowders. The translational diffusion coefficients are consistent with the association times (Table 3). A series of BD simulations show that intermolecular attractions are the major factors that might account for the large increase in association time of small molecules binding to HIVp. Notably, our ligands have only small partial charges in each bead and are neutral; thus, without long-range and strong electrostatic attractions, nonspecific van der Waals attractions contribute significantly to crowding effects. More studies with various sizes and charges of ligands are needed to study the crowding effects.

CONCLUSIONS

Ligand–protein complex formation can be viewed as a multi-step process. The initial step is a collision between two molecules in a solution. In the absence of electrostatic steering forces, this step is driven mainly by translational diffusion, and a crowded environment slows down the small-molecule diffusion process by 2- to 4-fold in our simulations. The second step involves several intermediates, before the third step, the final formation of the bound complex. This work investigated the molecular recognition processes that connect the initial and final steps of the binding process. We found that the ligand–protein interactions alter the flap dynamics from slow to fast gating. The interactions induced the flaps to open more frequently, although not sizably. Therefore, the gating effects became less significant in our work. Three ligands bind to HIVp with similar association times, but saquinavir and ritonavir fluctuated considerably in the bound state, so the two drugs may need a longer time to adjust their contact interface with HIVp. Even though our CG model cannot reveal all details, including water rearrangement, during ligand binding, the model provides valuable insights into the binding processes and guides us to further studies using detailed atomistic models. CGBD simulations can also simulate ligand–protein associations in the physiological ligand concentration.

ASSOCIATED CONTENT

S Supporting Information. Additional data, coarse-grained parameters, and potential functions are included. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: 951-827-7263; E-mail: chiaenc@ucr.edu.

ACKNOWLEDGMENT

We thank Dr. Joanna Trylska for valuable suggestions and discussion. This research was supported by the U.S. National Science Foundation (MCB-0919586). Additional support from the University of California, Riverside, Computer and Communications, and ShaRCS, the University of California Shared Research Computing Services Cluster, which is technically supported by multiple U.C. information technology divisions and managed by the University of California, Office of the President, is gratefully acknowledged.

REFERENCES

- (1) Mollerup, J. M. *Chem. Eng. Technol.* **2008**, *31*, 864–874.
- (2) Janin, J. *Prog. Biophys. Mol. Biol.* **1995**, *64*, 145–166.
- (3) Lundberg, S.; Backman, L. *Methods Enzymol.* **1994**, *228*, 241–254.
- (4) Mereghetti, P.; Kokh, D.; McCammon, J. A.; Wade, R. C. *BMC Biophys.* **2011**, *4*, 2.
- (5) Gabdoulline, R. R.; Wade, R. C. *Methods* **1998**, *14*, 329–341.
- (6) Smoluchowski, M. V. *Z. Phys. Chem.* **1917**, *92*, 129–168.
- (7) Schreiber, G.; Haran, G.; Zhou, H. X. *Chem. Rev.* **2009**, *109*, 839–860.
- (8) Furfine, E. S.; Dsouza, E.; Ingold, K. J.; Leban, J. J.; Spector, T.; Porter, D. J. T. *Biochemistry* **1992**, *31*, 7886–7891.
- (9) Luty, B. A.; McCammon, J. A. *Mol. Simul.* **1993**, *10*, 61–65.
- (10) Horn, J. R.; Sosnick, T. R.; Kossiakoff, A. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 2559–2564.
- (11) Kozer, N.; Kuttner, Y. Y.; Haran, G.; Schreiber, G. *Biophys. J.* **2007**, *92*, 2139–2149.
- (12) Wang, J.; Zhang, K.; Lu, H. Y.; Wang, E. *Biophys. J.* **2006**, *91*, 866–872.
- (13) McCammon, J. A.; Northrup, S. H. *Nature* **1981**, *293*, 316–317.
- (14) Northrup, S. H.; McCammon, J. A. *J. Am. Chem. Soc.* **1984**, *106*, 930–934.
- (15) Szabo, A.; Shoup, D.; Northrup, S. H.; McCammon, J. A. *J. Chem. Phys.* **1982**, *77*, 4484–4493.
- (16) Ellis, R. J. *Curr. Opin. Struct. Biol.* **2001**, *11*, 114–119.
- (17) Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Heimbach, J. C.; Dixon, R. A. F.; Scolnick, E. M.; Sigal, I. S. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686–4690.
- (18) Krausslich, H. G.; Ingraham, R. H.; Skoog, M. T.; Wimmer, E.; Pallai, P. V.; Carter, C. A. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 807–811.
- (19) Chang, C. E.; Shen, T.; Trylska, J.; Tozzini, V.; McCammon, J. A. *Biophys. J.* **2006**, *90*, 3880–3885.
- (20) Markgren, P. O.; Lindgren, M. T.; Gertow, K.; Karlsson, R.; Hamalainen, M.; Danielson, U. H. *Anal. Biochem.* **2001**, *291*, 207–218.
- (21) Shuman, C. F.; Markgren, P. O.; Hamalainen, M.; Danielson, U. H. *Antiviral Res.* **2003**, *58*, 235–242.
- (22) Spinelli, S.; Liu, Q. Z.; Alzari, P. M.; Hirel, P. H.; Poljak, R. J. *Biochimie* **1991**, *73*, 1391–1396.
- (23) Ang, W. H.; Parker, L. J.; De Luca, A.; Juillerat-Jeanneret, L.; Morton, C. J.; Lo Bello, M.; Parker, M. W.; Dyson, P. J. *Angew. Chem., Int. Ed.* **2009**, *48*, 3854–3857.
- (24) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C. H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Ericksonviitanen, S. *Science* **1994**, *263*, 380–384.
- (25) Clemente, J. C.; Stow, L. R.; Janka, L. K.; Jeung, J. A.; Desai, K. A.; Govindasamy, L.; Agbandje-McKenna, M.; McKenna, R.; Goodenow, M. M.; Dunn, B. M., In vivo, kinetic, and structural analysis of the development of ritonavir resistance. DOI: 10.2210/pdb2b60/pdb.
- (26) King, N. M.; Prabu-Jeyabalan, M.; Bandaranayake, R. M.; Nalam, M. N. L.; Ozen, A.; Hailloglu, T.; Schiffer, C. A., Extreme entropy-enthalpy compensation correlates with a drug resistant variant of HIV-1 protease. DOI: 10.2210/pdb3ekq/pdb.
- (27) Gopal, S. M.; Mukherjee, S.; Cheng, Y. M.; Feig, M. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1266–1281.

- (28) Chu, J. W.; Izvekov, S.; Voth, G. A. *Mol. Simul.* **2006**, *32*, 211–218.
- (29) Muller-Plathe, F. *Chemphyschem* **2002**, *3*, 754–769.
- (30) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- (31) Trylska, J.; Tozzini, V.; McCammon, J. A. *Biophys. J.* **2005**, *89*, 1455–1463.
- (32) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123–128.
- (33) Tozzini, V.; Trylska, J.; Chang, C. E.; McCammon, J. A. *J. Struct. Biol.* **2007**, *157*, 606–615.
- (34) Chang, C. E. A.; Trylska, J.; Tozzini, V.; McCammon, J. A. *Chem. Biol. Drug Design* **2007**, *69*, 5–13.
- (35) Reith, D.; Putz, M.; Muller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624–1636.
- (36) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352–1360.
- (37) Davis, M. E.; Madura, J. D.; Luty, B. A.; McCammon, J. A. *Comput. Phys. Commun.* **1991**, *62*, 187–197.
- (38) Geyer, T.; Winter, U. *J. Chem. Phys.* **2009**, 130.
- (39) Northrup, S. H.; Allison, S. A.; McCammon, J. A. *J. Chem. Phys.* **1984**, *80*, 1517–1526.
- (40) Judd, D. A.; Nettles, J. H.; Nevins, N.; Snyder, J. P.; Liotta, D. C.; Tang, J.; Ermolieff, J.; Schinazi, R. F.; Hill, C. L. *J. Am. Chem. Soc.* **2001**, *123*, 886–897.
- (41) Brynda, J.; Rezacova, P.; Fabry, M.; Horejsi, M.; Stouracova, R.; Soucek, M.; Hradilek, M.; Konvalinka, J.; Sedlacek, J. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2004**, *60*, 1943–1948.
- (42) Minh, D. D. L.; Chang, C. E.; Trylska, J.; Tozzini, V.; McCammon, J. A. *J. Am. Chem. Soc.* **2006**, *128*, 6006–6007.
- (43) Senapati, S.; Wong, C. F.; McCammon, J. A. *J. Chem. Phys.* **2004**, *121*, 7896–7900.
- (44) Li, D. C.; Liu, M. S.; Ji, B. H.; Hwang, K.; Huang, Y. G. *J. Chem. Phys.* **2009**, 130.
- (45) Marashi, S. A.; Behrouzi, R. *Biochem. Biophys. Res. Commun.* **2005**, *333*, 1–4.
- (46) Długosz, M.; Trylska, J. *BMC Biophys.* **2011**, 4.
- (47) Qin, S. B.; Minh, D. D. L.; McCammon, J. A.; Zhou, H. X. *J. Phys. Chem. Lett.* **2010**, *1*, 107–110.
- (48) Zaid, I. M.; Lomholt, M. A.; Metzler, R. *Biophys. J.* **2009**, *97*, 710–721.
- (49) Cheung, M. S.; Klimov, D.; Thirumalai, D. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 4753–4758.
- (50) Stagg, L.; Christiansen, A.; Wittung-Stafshede, P. *J. Am. Chem. Soc.* **2011**, *133*, 646–648.
- (51) Dix, J. A.; Verkman, A. S. *Annu. Rev. Biophys.* **2008**, *37*, 247–263.
- (52) McGuffee, S. R.; Elcock, A. H. *PLoS Comp. Biol.* **2010**, *6*.
- (53) Ando, T.; Skolnick, J. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18457–18462.
- (54) Wiczorek, G.; Zielenkiewicz, P. *Biophys. J.* **2008**, *95*, S030–S036.
- (55) Markgren, P. O.; Schaal, W.; Hamalainen, M.; Karlen, A.; Hallberg, A.; Samuelsson, B.; Danielson, U. H. *J. Med. Chem.* **2002**, *45*, 5430–5439.

Structures, Energetics, and IR Spectra of Monohydrated Inorganic Acids: Ab initio and DFT Study

Maciej Kołaski,^{†,‡} Aleksey A. Zakharenko,[†] S. Karthikeyan,[†] and Kwang S. Kim^{*,†}[†]Center for Superfunctional Materials, Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, 790-784 Pohang, South Korea Supporting Information

ABSTRACT: We carried out extensive calculations of diverse inorganic acids interacting with a single water molecule, through a detailed analysis of many possible conformations. The optimized structures were obtained by using density functional theory (DFT) and the second order Møller–Plesset perturbation theory (MP2). For the most stable conformers, we calculated the interaction energies at the complete basis set (CBS) limit using coupled cluster theory with single, double, and perturbative triple excitations [CCSD(T)]. The –OH stretching harmonic and anharmonic frequencies are provided as fingerprints of characteristic conformers. The zero-point energy (ZPE) uncorrected/corrected ($\Delta E_c/\Delta E_0$) interaction energies and the enthalpies/free energies ($\Delta H_r/\Delta G_r$) at room temperature and 1 bar) are reported. Various comparisons are made between many diverse inorganic acids (H_mXO_n where $X = B/N/P/Cl/Br/I$, $m = 1-3$, and $n = 0-4$) as well as other simple inorganic acids. In many cases, we find that the dispersion-driven van der Waals interactions between X in inorganic acid molecules and O in water molecules as well as the $X^+ \cdots O^-$ electrostatic interactions are important.

INTRODUCTION

The hydration phenomena of diverse molecular systems including cations,¹ anions,² simple acids³/bases,⁴ and salts⁵ have been reported. However, the structures, spectra, and thermochemical data for hydration of a large family of inorganic acids with boron, nitrogen, phosphorus, and sulfur atoms as well as halide atoms are still scarce.⁶ Thus, we are interested in studying the hydration of such inorganic acids: HBO_2/HBO_3 , HNO_2/HNO_3 , H_3PO_3/H_3PO_4 , H_2SO_3/H_2SO_4 , H_2S , H_2Se , HCN , HF , $HClO_{n=0-4}$, $HBrO_{n=0-4}$, and $HIO_{n=0-4}$. However, due to high complexity of the hydration/dehydration phenomena which requires extensive studies, the present aim is not to study such complex phenomena, but we have tried to investigate the monohydrated inorganic acids as the first step toward understanding hydration of the above 27 inorganic acids. The information from various monohydrated inorganic acids would be useful in studying both protonated acid cations and deprotonated acid anions.

We begin with brief introduction of various inorganic acids studied here. HCl , HBr , HI , and HNO_3 are strong monoprotic acids, and H_2SO_4 is a biprotic strong acid. H_3BO_3 and H_3PO_4 are weak triprotic acids. Hypochlorous ($HClO$), hypobromous ($HBrO$), and hypoiodous (HIO) acids are the simplest examples of weak acids. Chlorous ($HClO_2$), bromous ($HBrO_2$), and iodous (HIO_2) acids are also relatively weak in comparison with other inorganic acids. Chloric acid ($HClO_3$) is known as a strong acid and a very efficient oxidizing agent. Iodic acid (HIO_3) exists as a white solid and is insoluble in water, unlike chloric and bromic acids. Perchloric acid ($HClO_4$) is a strong acid which completely dissociates in an aqueous solution. Its strength is comparable to sulfuric acid (H_2SO_4) and nitric acid (HNO_3). Perbromic acid ($HBrO_4$) is a strong acid and an important oxidizing agent in many chemical reactions. Periodic acid (HIO_4) is widely employed in organic chemistry for structural analysis. Periodic acid

can cleave a vicinal diol into two aldehyde or ketone fragments, which is useful in determining the structure of carbohydrates.

The nitric acid (HNO_3)⁷⁻¹⁷ and sulfuric acid (H_2SO_4)¹⁸⁻²⁵ were widely studied theoretically. Owing to experimental difficulties, less attention is paid to sulfurous acid [H_2SO_3 or $HS(OH)O_2$],^{26,27} while little is known about sulfonic acid [$(HO)_2SO$] which has the same chemical formula. The sulfur-containing acid with the simplest chemical formula H_2S (hydrogen sulfide) is a covalent hydride related structurally to the water molecule, since oxygen and sulfur belong to the same group of the periodic table of elements.^{28,29} Hydrogen selenide (H_2Se) is the simplest hydride of selenium, a colorless flammable gas under standard conditions, and soluble in water. The physicochemical properties of H_2S and H_2Se are similar.

The phosphoric acid (H_3PO_4) is used as the electrolyte in fuel cells, and the phosphorous acid (H_3PO_3 ; $HP(OH)_2O$) is commonly used as oxoacids of phosphorus.³⁰⁻³⁴ As in H_2SO_3 , the isomer $P(OH)_3$ is not known well, like sulfonic acid $(HO)_2SO$. Phosphoric acid may be used as a “rust converter”, by direct application to rusted iron, steel tools, or surfaces. The hydrogen cyanide (HCN) is a valuable precursor to many chemical compounds ranging from polymers to pharmaceuticals. Hydrogen cyanide is a colorless, very poisonous, and highly volatile liquid that boils slightly above room temperature at 26 °C.³⁵⁻³⁷ It is important to note that HCN is one of the simplest cyanide systems which can act as both proton donor and acceptor.

In this work, we compare the interaction energies, H-bond lengths (r_{OH}), natural bond orbital (NBO) charges (q), and –OH IR (infrared) stretching harmonic/anharmonic vibrational frequencies (ν_s) using DFT and extensive high-level ab initio

Received: August 2, 2010

Published: August 25, 2011

calculations of diverse inorganic acids interacting with a single water molecule.

COMPUTATIONAL DETAILS

We have carried out geometry optimization, harmonic/anharmonic frequency analysis, and calculated interaction energies for many different monohydrated inorganic acids. Geometry optimization and harmonic frequency calculations were done at the DFT level of theory employing Becke's three-parameter exchange potential and the Lee–Yang–Parr correlation functional (B3LYP) as well as second order Møller–Plesset perturbation (MP2) theory. The anharmonic frequencies were obtained at the MP2 level of theory. Calculations were carried out by using the Gaussian 03 suite of programs.³⁸ The zero-point energy correction is only the second order for the anharmonic terms, and so this anharmonic correction is insignificant in general. The thermal energy corrections are calculated based on rigid rotor and harmonic oscillator approximation. Throughout the present work, all the atoms were treated with the aug-cc-pVDZ basis set (which we have abbreviated as aVDZ) and the aug-cc-pVTZ basis set (which will be abbreviated as aVTZ). Since the aVDZ/aVTZ basis set is not available for iodine, the CRENL ECP basis set was employed. This basis set uses relativistic effective core potentials and an extended valence basis set. Since only one CRENL ECP basis set alone cannot be exploited for the basis set extrapolation to estimate the complete basis set (CBS) limit energy (even though all atoms are treated with aVDZ and aVTZ basis sets), the reported CBS energies for iodine-containing clusters are approximate in that the extrapolation related to the I atom is not properly taken into account. All optimized structures were drawn with the Posmol package.³⁹ To obtain more reliable results, we carried out single point coupled cluster theory with single, double, and perturbative triple excitations [CCSD(T)] calculations for the most stable conformers of monohydrated acids. To estimate the CCSD(T)/CBS interaction energies, we obtain the CBS limit values for the MP2 interaction energies using the extrapolation method exploiting that the basis set error in the electron correlation energy is proportional to N^{-3} for the aug-cc-pVNZ (which we have abbreviated as aVNZ) basis set (aVDZ for $N = 2$, aVTZ for $N = 3$).^{40,41} Given that the difference in interaction energy between MP2/aVNZ and CCSD(T)/aVNZ does not differ significantly with increasing size of the basis set, the CCSD(T)/CBS interaction energies are evaluated from the MP2/CBS ones by using the difference between CCSD(T)/aVDZ and MP2/aVDZ interaction energies.^{41,42} The CCSD(T) calculations were carried out by using the MOLPRO suite of programs.⁴³ Then, these interaction energies (ΔE_e) are used to evaluate the zero-point energy (ZPE) corrected interaction energies (ΔE_0) and the enthalpy and Gibbs free energies at room temperature (298.15 K) and 1 bar, by using the MP2/aVDZ ZPEs and thermal energies.

All the reported structures are at the local or global minima without imaginary frequencies. All structures found for one type of acid...water clusters were attempted for all similar monohydrated inorganic acids. To distinguish different conformers, we use the following notation: A_Y/D_Z denotes the role of proton acceptor (A: O)/donor (D: H) by the water molecule, while the subscript " $_{Y/Z}$ " denotes the atom of each acid interacting with the O/H atom of H_2O . Since $A_{H(O)}$ is the most common case, it will be simply denoted as A, which is distinguished from $A_{H(X)}$ to be simply denoted as A_H , where X is the central atom in inorganic acid. Here, subscript H(O)/H(X) indicates that in inorganic acid, the

H atom is bonded to the O/X atom, while O(H)/O(X) indicates that the O atom is bonded to the H/X atom. The distance between the two H-bonded O atoms (one in H_2O and the other in acid) is denoted as r_{OO} . The $D_{O(H)}$ structure, where a H atom in H_2O interacts with the $-O(H)$ group in the acid, is simply denoted as D_{OH} , in contrast to the D_O structure, where a H atom in H_2O interacts with a double-bond oxygen ($O=$) in the acid. The H-bond distance between an O atom in an acid molecule and a H atom in H_2O is denoted as r_{OH} .

There are a few geometrical isomers showing the difference only in H orientations. The energy differences and barriers between these isomers due to simple H orientations are not large. These structures are floppy, and so we have discussed mainly the low-lying energy structures. The qualitative trends are not likely to change by including the facile motions of the water molecule. In this study we focus our attention on the minimum-energy structures which are highly stable. Thus, the most stable conformer among topologically different isomers for each acid is marked in bold, and in this case the CCSD(T)/CBS interaction energies and thermodynamic quantities are evaluated.

In many cases, the central atom of an acid molecule (X) interacts with the oxygen (O) atom of a water molecule through the dispersion-driven van der Waals interaction. The van der Waals radii for N, O, P, S, Cl, Br, and I are 1.55, 1.52, 1.80, 1.80, 1.75, 1.85, and 1.98 Å, respectively.⁴⁴ If the distance between X and O (r_{XO}) is close to the sum of van der Waals radii of X and O ($r_{XO} \approx r_X + r_O$), then the notation W_{XO} is employed to denote the van der Waals interaction. In some cases, the distance between X and O is significantly shorter when the electrostatic interaction between the highly positively charged X and the negatively charged O is much stronger than the dispersion-driven van der Waals interaction. In such cases, the notation E_{XO} is used to denote the electrostatic interaction between the positively charged acid X atom and the negatively charged water O atom. We note that formation of monohydrated inorganic acids is determined by hydrogen bonding, dispersion-driven van der Waals interaction, and electrostatic interaction between the positively charged acid X atom (if any) and the negatively charged water O atom. Thus, our discussion on the r_{XO} distances is based on the van der Waals radii and the NBO charges localized on X and O. The NBO charges (q in au) of the X/Y atom will be denoted as $q_{X/Y}$. To support this analysis, we provide a more in-depth analysis. We carried out a series of symmetry adapted perturbation theory (SAPT) calculations⁴⁵ based on DFT^{46,47} of monohydrated $HXO_{n=0,2}$ complexes, since the X atom in HX is negatively charged, and that in HXO is weakly charged, while that in HXO_2 or HXO_3 is positively charged. Symmetry-adapted perturbation theory (SAPT)-DFT, which takes into account the dispersion energy, provides a detailed description of molecular interactions in clusters, which are decomposed into electrostatic, induction, dispersion, and exchange parts. However, the decomposition between two specific atomic sites of molecules is not possible. Thus, the analysis is qualitatively investigated. The interaction energy is decomposed into electrostatic (E_{es}), induction (E_{ind}), dispersion (E_{disp}), and exchange repulsion (E_{exch}). However, the exchange–induction term ($E_{ind-exch}$) and exchange–dispersion term ($E_{disp-exch}$) can often be added to E_{ind} and E_{disp} , respectively, to form the effective induction ($E_{ind^*} = E_{ind} + E_{ind-exch}$) and the effective dispersion ($E_{disp^*} = E_{disp} + E_{disp-exch}$), respectively, while the two terms are extracted from E_{exch} to form the effective exchange term ($E_{exch^*} = E_{exch} - (E_{ind-exch} + E_{disp-exch})$) as described previously.^{48,49}

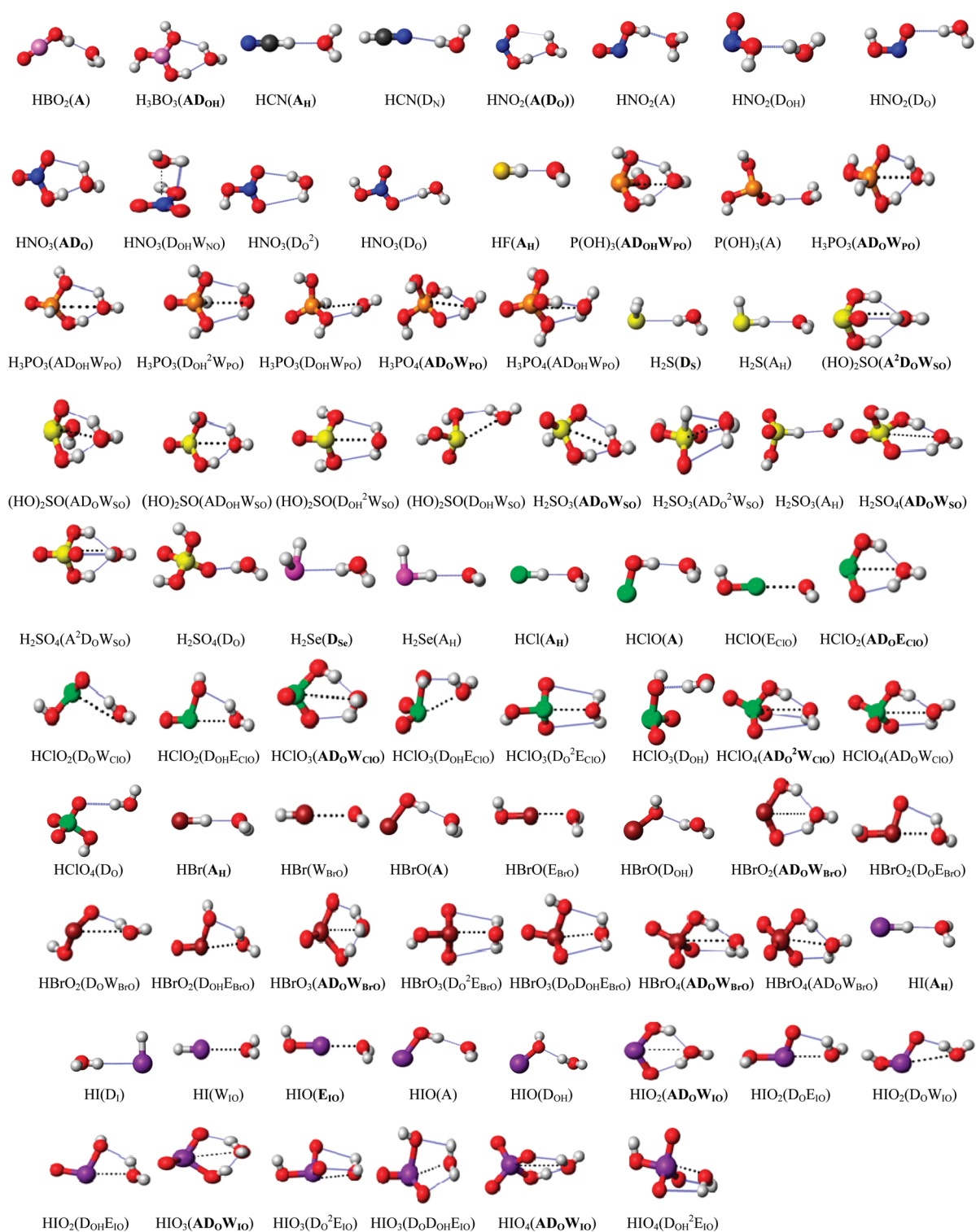


Figure 1. MP2/aVDZ optimized structures of monohydrated inorganic acids. The lowest energy structure for each chemical species is denoted in bold. Dashed lines represent hydrogen bonds. Dotted bold lines denote van der Waals or electrostatic interactions.

RESULTS AND DISCUSSION

Figure 1 shows various stable structures at the MP2/aVDZ level of theory. Tables 1 and 2 list MP2/aVDZ and B3LYP/aVDZ energies (ZPE-uncorrected interaction energies: ΔE_c) and structural parameters. The ZPE-corrected interaction energies (ΔE_0) and enthalpies/free energies ($\Delta H_T/\Delta G_T$ at room

temperature and 1 bar) are in Supporting Information (Tables S1, S2). The geometrical parameters include the H-bond distances between acids and a water molecule (r_{OO} , r_{OH}) and the van der Waals interaction distance between the X atom of the acid and the O atom in the water molecule (r_{XO}). For the most stable structures we calculated the NBO charges at the

Table 1. MP2(B3LYP)/aVDZ Interaction Energies (ΔE_e in kJ/mol) and Selected Interatomic Distances for Various Conformers of Inorganic Acid \cdots Water Complexes^a

acid \cdots water	conformer	$-\Delta E_e$	r_{OO}	r_{OH}	r_{XO}
HBO ₂ \cdots H ₂ O	A	42.7(40.9)	2.71(2.70)	1.72(1.71)	
H ₃ BO ₃ \cdots H ₂ O	AD _{OH}	37.8(31.4)	2.82(2.83)	2.02(2.03)	
HCN \cdots H ₂ O	A _H	23.7(21.3)		2.05(2.04)	
	D _N	18.5(15.2)			
HNO ₂ \cdots H ₂ O	A(D _O)	32.3(29.3)	2.99(3.01)	2.53(2.54)	3.39(3.38)
	A	30.9(27.7)	2.79(2.78)	1.80(1.80)	3.25(3.34)
	D _{OH}	15.9(–)	2.96(–)	2.05(–)	3.59(–)
	D _O	10.3(7.6)	3.10(3.13)	2.15(2.17)	
HNO ₃ \cdots H ₂ O	AD _O	43.1(40.0)	2.91(2.91)	2.39(2.39)	3.26(3.26)
	D _{OH} W _{NO}	16.4(–)	2.87(–)	2.83(–)	2.83(–)
	D _O ²	12.3(8.2)	3.16(3.17)	2.22(2.20)	
	D _O	11.8(8.0)	3.61(3.98)	2.15(2.18)	3.61(3.98)
P(OH) ₃ \cdots H ₂ O	AD _{OH} W _{PO}	41.1(31.7)	2.91(2.91)	2.07(2.07)	3.44(3.38)
	A	18.2(13.9)	2.82(2.83)	1.84(1.85)	3.74(3.83)
H ₃ PO ₃ \cdots H ₂ O	AD _O W _{PO}	54.8(48.7)	2.78(2.77)	1.92(1.91)	3.25(3.25)
	AD _{OH} W _{PO}	34.9(28.1)	2.86(2.91)	2.18(2.22)	3.35(3.46)
	D _{OH} ² W _{PO}	19.5(10.9)	3.01(3.13)	2.43(2.52)	3.24(3.39)
	D _{OH} W _{PO}	18.7(12.3)	2.90(2.95)	2.12(2.13)	3.37(3.52)
H ₃ PO ₄ \cdots H ₂ O	AD _O W _{PO}	54.9(48.3)	2.79(2.79)	1.94(1.93)	3.24(3.24)
	AD _{OH} W _{PO}	40.0(32.3)	2.96(2.98)	2.25(2.30)	3.36(3.44)
H ₂ S \cdots H ₂ O	D _S	14.4(11.0)			
	A _H	13.4(9.5)		2.19(2.22)	
(HO) ₂ SO \cdots H ₂ O	A ² D _O W _{SO}	49.0(39.2)	2.73(2.73)	1.99(1.98)	3.08(3.10)
	AD _O W _{SO}	45.7(38.4)	2.84(2.82)	2.03(2.01)	3.27(3.31)
	AD _{OH} W _{SO}	37.5(28.6)	2.88(2.96)	2.19(2.29)	3.22(3.22)
	D _{OH} ² W _{SO}	18.0(9.0)	3.00(3.14)	2.45(2.51)	3.08(3.30)
	D _{OH} W _{SO}	15.6(10.4)	2.91(3.02)	2.13(2.06)	3.21(4.16)
H ₂ SO ₃ \cdots H ₂ O	AD _O W _{SO}	52.3(46.4)	2.83(2.83)	2.09(2.09)	3.34(3.34)
	AD _O ² W _{SO}	28.4(–)	3.14(–)	2.80(–)	2.93(–)
	A _H	25.4(22.0)		2.00(1.97)	
H ₂ SO ₄ \cdots H ₂ O	AD _O W _{SO}	53.3(46.4)	2.88(2.88)	2.17(2.17)	3.30(3.33)
	A ² D _O W _{SO}	47.6(37.1)	2.77(2.79)	2.18(2.20)	3.04(3.08)
	D _O	11.5(6.5)	3.03(3.06)	2.08(2.10)	
H ₂ Se \cdots H ₂ O	D _{Se}	14.9(10.0)			
	A _H	12.1(7.3)		2.23(2.30)	

^aH-bond: r_{OH} ; distance between two H-bonded O atoms: r_{OO} , van der Waals or electrostatic interaction: r_{XO} in Å, r_{XO} denotes the X(acid) \cdots OH₂ distance; r_{OH} denotes either H(acid) \cdots OH₂ or O(acid) \cdots HOH distance; and r_{OO} denotes the O(acid) \cdots OH₂ distance. The (–) means that these structures do not exist at the B3LYP level of theory (lack of dispersion energy). The most stable structures are marked in bold.

MP2/aVDZ level of theory. Since the NBO charges localized on water molecule are almost identical in all the cases, we report in Table 3 the NBO charges ($q_{H(O)}$, $q_{O(H)}$, q_X) of the atoms in inorganic acids involved in H-bonding (O/H) formation and the van der Waals interaction (X–O). For the most stable conformers we also calculated ΔE_e , ΔE_0 , ΔH_r , and ΔG_r at the CCSD(T)/CBS level of theory including basis set superposition error (BSSE)-corrected MP2/aVDZ, MP2/aVTZ, and CCSD(T)/aVDZ interaction energies (Table 4). In the majority of cases, B3LYP/aVDZ approach underestimates the binding energies (the absolute value of the interaction energies) as compared to the CCSD(T)/CBS results. However, it provides the proper order of conformers. Since B3LYP/aVDZ calculations are not computationally demanding, B3LYP is a good starting point for more accurate calculations by searching for many possible conformations.

In some cases, to find new structures we carry out MP2 geometry optimizations. If the formation of clusters is driven by dispersion interactions, then the B3LYP method does not allow us to identify such isomers. The MP2/aVDZ approach gives reasonable values of the interaction energies, since the energy difference between CCSD(T)/CBS and MP2/aVDZ in the present systems is small in most cases, except for small iodine-containing clusters, such as HI and HIO, for which the aVDZ is not sufficient for iodine. Thus, for energy comparison between different isomers to find the most stable structures, our discussion will be based on the MP2/aVDZ ΔE_0 , unless otherwise specified.

Monohydrated Inorganic Acid Complexes with B, C, N, F, P, S, Se, F, Cl, Br, and I. In the case of monohydrated metaboric acid (HBO₂), only one stable structure (A) is found. The monohydrated orthoboric acid (H₃BO₃) also forms one stable

Table 2. MP2(B3LYP)/aVDZ Interaction Energies (ΔE_c in kJ/mol) and Selected Interatomic Distances for Various Conformers of Halogen Containing Acid \cdots Water Complexes^a

acid \cdots water	conformer	$-\Delta E_c$	r_{OO}	r_{OH}	r_{XO}
HF \cdots H ₂ O	A _H	37.8(37.7)			
HCl \cdots H ₂ O	A _H	25.9(23.1)			
HClO \cdots H ₂ O	A	33.6(30.0)	2.78(2.78)	1.79(1.79)	3.51(3.59)
	E _{ClO}	13.6(10.2)			2.77(2.76)
HClO ₂ \cdots H ₂ O	AD _O E _{ClO}	46.1(41.3)	2.84(2.85)	2.03(2.02)	2.99(3.05)
	D _O W _{ClO}	22.3(18.0)	2.87(2.91)	1.97(1.96)	3.28(3.57)
	D _{OH} E _{ClO}	21.6(11.8)	2.87(2.95)	2.17(2.29)	2.94(3.11)
HClO ₃ \cdots H ₂ O	AD _O W _{ClO}	40.8(39.3)	2.96(2.92)	2.28(2.22)	3.27(3.32)
	D _{OH} E _{ClO}	27.6(14.0)	2.81(2.93)	2.03(2.42)	2.91(2.97)
	D _O ² E _{ClO}	22.7(15.2)	3.15(3.17)	3.15(3.05)	2.79(2.86)
	D _{OH}	18.50(–)	2.94(–)	2.01(–)	4.08(–)
HClO ₄ \cdots H ₂ O	AD _O ² W _{ClO}	49.3(44.4)	3.09(3.19)	2.56(2.75)	3.28(3.40)
	AD _O W _{ClO}	47.0(42.7)	3.00(3.19)	2.45(2.75)	3.31(3.41)
	D _O	10.4(5.3)	3.21(3.28)	2.29(2.31)	4.01(4.18)
HBr \cdots H ₂ O	A _H	22.8(19.0)			
	W _{BrO}	8.2(3.8)			3.11(3.23)
HBrO \cdots H ₂ O	A	32.3(28.0)	2.79(2.80)	1.81(1.81)	3.63(3.71)
	E _{BrO}	21.3(16.6)			2.75(2.73)
	D _{OH}	18.3(–)	2.91(–)	2.02(–)	3.98(–)
HBrO ₂ \cdots H ₂ O	AD _O W _{BrO}	48.6(41.3)	2.83(2.84)	1.98(1.99)	3.07(3.15)
	D _O E _{BrO}	28.7(21.2)	2.94(2.94)	2.34(2.29)	2.83(2.85)
	D _O W _{BrO}	27.3(20.8)	2.84(2.88)	1.93(1.93)	3.33(3.62)
	D _{OH} E _{BrO}	26.9(15.7)	2.86(2.93)	2.22(2.30)	2.92(3.03)
HBrO ₃ \cdots H ₂ O	AD _O W _{BrO}	48.0(43.1)	2.87(2.86)	2.10(2.07)	3.33(3.38)
	D _O ² E _{BrO}	33.3(24.6)	3.11(3.13)	2.90(2.91)	2.76(2.78)
	D _O D _{OH} E _{BrO}	33.3(21.4)	2.97(3.02)	2.64(2.75)	2.82(2.86)
HBrO ₄ \cdots H ₂ O	AD _O W _{BrO}	54.4(47.2)	3.02(3.02)	2.38(2.41)	3.29(3.43)
	AD _O W _{BrO}	52.3(46.4)	2.93(2.91)	2.29(2.27)	3.37(3.47)
HI \cdots H ₂ O	A _H	17.3(12.2)			
	D _I	16.9(7.6)			
	W _{IO}	11.8(9.6)			3.18(3.16)
HIO \cdots H ₂ O	E _{IO}	30.5(28.5)			2.76(2.75)
	A	29.9(24.9)	2.83(2.84)	1.84(1.86)	3.76(3.84)
	D _{OH}	25.4(18.5)	2.88(2.90)	1.93(1.98)	3.88(4.35)
HIO ₂ \cdots H ₂ O	AD _O W _{IO}	54.6(41.8)	2.78(2.82)	1.89(1.93)	3.16(3.24)
	D _O E _{IO}	42.2(35.1)	2.82(2.80)	2.12(2.08)	2.78(2.78)
	D _O W _{IO}	37.5(28.7)	2.79(2.82)	1.84(1.87)	3.41(3.62)
	D _{OH} E _{IO}	35.7(24.3)	2.80(2.86)	2.11(2.18)	2.94(3.02)
HIO ₃ \cdots H ₂ O	AD _O W _{IO}	54.7(47.6)	2.81(2.80)	1.96(1.98)	3.42(3.50)
	D _O ² E _{IO}	45.7(39.1)	3.02(3.03)	2.67(2.69)	2.70(2.69)
	D _O D _{OH} E _{IO}	42.9(32.4)	2.89(2.94)	2.49(2.50)	2.79(2.82)
HIO ₄ \cdots H ₂ O	AD _O W _{IO}	56.3(49.8)	2.89(2.92)	2.14(2.18)	3.52(3.55)
	D _{OH} ² E _{IO}	39.5(33.4)	2.86(2.86)	2.60(2.61)	2.59(2.59)

^a H-bond: r_{OH} ; distance between two H-bonded O atoms: r_{OO} ; van der Waals or electrostatic interaction: r_{XO} in Å; r_{XO} denotes the X(acid) \cdots OH₂ distance; r_{OH} denotes either H(acid) \cdots OH₂ or O(acid) \cdots HOH distance; r_{OO} denotes the O(acid) \cdots OH₂ distance. The (–) means that these structures do not exist at the B3LYP level of theory (lack of dispersion energy). The most stable structures are marked in bold.

conformer (AD_{OH}). The H-bond length for HBO₂ ($r_{OH} = 1.72$ Å) is shorter than that for H₃BO₃ ($r_{OH} = 2.02$ Å).

The monohydrated hydrogen cyanide (HCN) forms two different stable structures. The most stable structure is A_H, which is ~ 6.2 kJ/mol more stable than the D_N structure. In the case of D_N conformer, the structure of $-N \cdots HO-$ motif is slightly bent ($\angle(\text{NHO}) = 176^\circ$).

In the case of monohydrated nitrous acid (HNO₂), the most stable structure A(D_O) has one full H-bond and a partial H-bond (which cannot be considered as a real H-bond because the r_{OH} distance is too long (>2.5 Å) and the bond angle ($\angle(\text{OHO})$ is too small ($\sim 109^\circ$)). Thus, this structure is denoted as A(D_O). It is slightly more stable than the A conformer (by ~ 1.3 kJ/mol). The comparison of H-bond lengths of the A(D_O) and D conformers

Table 3. NBO Charges Localized on the Lowest Energy Structure of Monohydrated Inorganic Acids at the MP2/aVDZ Level^a

acid	structure	q_X	$q_{H(O)}$	$q_{O(H)}$	$q_{O=}$	q_H
HBO ₂	A	1.46	0.57	-1.04	-1.02	
H ₃ BO ₃	AD _{OH}	1.53	0.55	-1.06		
HCN	A _H	0.14				0.25
HNO ₂	A(D _O)	0.51	0.53	-0.64	-0.43	
HNO ₂	A	0.45	0.54	-0.66	-0.35	
HNO ₃	AD _O	0.88	0.56	-0.58	-0.50	
HF	A _H	-0.63				0.60
P(OH) ₃	AD _{OH} W _{PO}	1.73	0.55	-1.12		
H ₃ PO ₃	AD _O W _{PO}	2.39	0.57	-1.11	-1.22	-0.11
H ₃ PO ₄	AD _O W _{PO}	2.80	0.57	-1.10	-1.23	
H ₂ S	D _S	-0.29				0.15
H ₂ S	A _H	-0.32				0.18
(HO) ₂ SO	A ² D _O W _{SO}	1.91	0.54	-0.95	-1.08	
(HO) ₂ SO	AD _O W _{SO}	1.90	0.56	-0.99	-1.06	
H ₂ SO ₃	AD _O W _{SO}	2.38	0.57	-0.98	-1.06	0.05
H ₂ SO ₄	AD _O W _{SO}	2.77	0.57	-0.96	-1.03	
H ₂ Se	D _{Se}	-0.14				0.08
HCl	A _H	-0.35				0.32
HClO	A	0.16	0.54	-0.72		
HClO ₂	AD _O E _{ClO}	1.05	0.55	-0.76	-0.85	
HClO ₃	AD _O W _{ClO}	1.93	0.55	-0.79	-0.87	
HClO ₄	AD _O ² W _{ClO}	2.61	0.56	-0.78	-0.83	
HClO ₄	AD _O W _{ClO}	2.61	0.56	-0.78	-0.84	
HBr	A _H	-0.26				0.24
HBrO	A	0.25	0.54	-0.81		
HBrO ₂	AD _O W _{BrO}	1.24	0.55	-0.84	-0.95	
HBrO ₃	AD _O W _{BrO}	2.26	0.56	-0.90	-0.99	
HBrO ₄	AD _O W _{BrO}	3.02	0.56	-0.88	-0.94	
HBrO ₄	AD _O W _{BrO}	3.02	0.57	-0.89	-0.94	
HI	A _H	-0.11				0.09
HI	D _I	-0.04				0.05
HIO	E _{IO}	0.46				
HIO	A	0.39	0.53	-0.94		
HIO ₂	AD _O W _{IO}	1.56	0.55	-0.99	-1.11	
HIO ₃	AD _O W _{IO}	2.82	0.57	-1.08	-1.18	
HIO ₄	AD _O W _{IO}	3.80	0.58	-1.08	-1.15	

^aNBO charges, q , in au. Subscript X denotes the central atom in the inorganic acid; H(O)/O(H) denotes a H/O atom bonded to an O/H atom. O= denotes an O atom which is double-bonded to X (if there are more such oxygen atoms, we report the most negative NBO charge); subscript H in q_H denotes an H atom directly attached to X. The NBO charges of the monohydrated HIO are not reported. The OH group of HIO is not involved in any hydrogen bond formation. All reported NBO charges are localized on the acid.

shows that in the case of D structures the H-bond lengths are slightly shorter (by ~ 0.4 Å).

In the case of monohydrated nitric acid (HNO₃), the most stable conformer is AD_O. This structure looks similar to the A(D_O) conformer of HNO₂ but different from it, because HNO₃ has two full H-bonds, while HNO₂ has only one full H-bond and another very weak H-bond. Thus, HNO₃ has a much stronger binding energy than HNO₂ (by ~ 10.5 kJ/mol). In the monohydrated HNO₃ the r_{NO} distance is shorter than that in the

monohydrated HNO₂. The charge q_N localized on the N atom in HNO₃ (0.88 au) is substantially larger than the q_N of HNO₂ (0.51 au). The D_{OH}W_{NO} conformer is less stable than AD_O, however, D_{OH}W_{NO} is additionally stabilized by van der Waals interaction between N and O. It is important to note that the formation of the D_{OH}W_{NO} structure is governed by dispersion energy. At the B3LYP level of theory, this structure cannot be found. The D_O² and D_O complexes are significantly less stable than the AD_O structure.

The monohydrated hydrogen fluoride (HF) and hydrogen chloride (HCl) form one stable conformer A_H.^{50,51} The interaction energy of ClH \cdots OH₂ is ~ 8.7 kJ/mol smaller in magnitude than that of FH \cdots OH₂. In the case of HBr and HI, both A_H and W_{XO} (X = Br, I) structures exist. It is worth noting that D_X (X = F, Cl, Br) conformers are transformed to different structures during geometry optimization (D_F to A_H, D_{Cl} to A_H, and D_{Br} to W_{BrO}). The W_{XO} structure is determined by van der Waals interaction between the heavy halogen (X) and the oxygen (O) atoms of the water molecule. In the case of monoprotic acids HX (X = F, Cl, Br, and I) the charge q_H decreases with the increasing atomic number of X. The q_H localized on HF, HCl, HBr, and HI are 0.60, 0.32, 0.24, and 0.09 au, respectively, consistent with the gradual decrease in interaction energy for the monohydration phenomenon. In the case of monohydrated HI, we find an additional D_I conformer. Its binding energy is very close to the A_H structure.

In the case of phosphorous acid (H₃PO₃), there are two different topological isomers: HP(OH)₂O and P(OH)₃. HP(OH)₂O is ~ 15.0 kJ/mol more stable than P(OH)₃. When the phosphorous acid (H₃PO₃; HP(OH)₂O) is monohydrated, the most stable structure is AD_OW_{PO}. The structures with a water molecule acting as a donor (D_{OH}²W_{PO} and D_{OH}W_{PO}) are significantly less stable. In the case of AD_OW_{PO} structure the H-bond length r_{OH} is 1.92 Å, while in the case of D_{OH}²W_{PO} structure, it is significantly longer (2.43 Å). When the P(OH)₃ acid is monohydrated, the most stable structure is AD_{OH}W_{PO}. Although both acids are isomers, their structures are quite different, since the P(OH)₃ acid does not contain —P=O group which excludes the formation of AD_O forms.

In the case of monohydrated phosphoric acid (H₃PO₄), the most stable structure is AD_OW_{PO}, while the AD_{OH}W_{PO} structure is less stable. The most stable structures of monohydrated H₃PO₃ and H₃PO₄ have similar interaction energies, because both conformers are determined by the AD_OW_{PO} motif.

For the monohydrated hydrogen sulfide (H₂S), the D_S structure is ~ 1 kJ/mol more stable in ΔE_e than the A_H structure. However, the relative interaction energy difference in ΔE_0 is only ~ 0.08 kJ/mol, then, the two structures are almost isoenergetic. The angle ($122 \pm 6^\circ$) between the O—(H) \cdots O axis and the bisecting axis of the out-of-plane H₂O molecule in the water dimer^{52–57} is in between the angle (92.5°) of the bisecting axis of the H₂S molecule with respect to the S \cdots (H)—O axis in the D_S structure and the angle (146°) of the bisecting axis of the H₂O molecule with respect to the O—(H) \cdots S axis in the A_H structure, because of a slightly stronger acidity for H₂S over H₂O.

The H₂SO₃ has two different isomers: sulfonic acid [HS(OH)O₂ (to be denoted simply as H₂SO₃)] and sulfurous acid [(HO)₂SO]. The sulfonic acid is ~ 4.6 kJ/mol more stable than the sulfurous acid. For the monohydrated sulfonic acid (H₂SO₃), the most stable conformer is AD_OW_{SO}. For the monohydrated sulfurous acid (HO)₂SO, the most stable structure A²D_OW_{SO} has three hydrogen bonds. The second lowest energy conformer

Table 4. BSSE-Corrected Interaction Energies (ΔE_e) at the MP2/aVTZ and CCSD(T)/aVDZ Levels of Theory Along with the CCSD(T)/CBS Thermodynamic Quantities Estimated with the MP2/aVDZ Thermal Energies^a

acid...water	structure	MP2	CCSD(T)	CCSD(T)/CBS				
		/aVTZ	/aVDZ	$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$	T_h
HBO ₂ ...H ₂ O	A	39.5	37.8	41.3	32.4	34.2	4.5	359
H ₃ BO ₃ ...H ₂ O	AD _{OH}	34.0	32.6	36.4	28.9	29.0	-8.1	245
HCN...H ₂ O	A _H	20.7	19.5	21.0	16.1	16.0	-7.7	236
HNO ₂ ...H ₂ O	A(D _O)	29.0	28.4	32.0	23.2	24.8	-9.1	225
HNO ₂ ...H ₂ O	A	29.0	28.4	30.8	22.2	22.9	-6.6	163
HNO ₃ ...H ₂ O	AD _O	39.1	39.7	43.9	34.9	36.7	0.8	298
HF...H ₂ O	A _H	34.9	33.4	36.2	24.8	29.3	-0.5	289
P(OH) ₃ ...H ₂ O	AD _{OH} W _{PO}	36.3	35.8	41.3	30.6	33.6	-5.8	246
H ₃ PO ₃ ...H ₂ O	AD _O W _{PO}	51.7	49.9	57.6	46.3	50.1	9.4	347
H ₃ PO ₄ ...H ₂ O	AD _O W _{PO}	51.2	50.1	58.0	47.1	50.6	11.3	360
H ₂ S...H ₂ O	D _S	11.8	10.1	11.8	5.4	5.9	-17.5	118
H ₂ S...H ₂ O	A _H	11.1	9.6	8.6	3.2	2.9	-19.1	129
(HO) ₂ SO...H ₂ O	A ² D _O W _{SO}	45.2	45.6	52.4	40.7	44.5	2.7	299
(HO) ₂ SO...H ₂ O	AD _O W _{SO}	42.0	44.1	50.7	39.6	43.0	2.6	288
H ₂ SO ₃ ...H ₂ O	AD _O W _{SO}	49.9	48.1	56.4	46.2	49.3	10.6	346
H ₂ SO ₄ ...H ₂ O	AD _O W _{SO}	49.8	48.3	56.0	46.2	49.0	10.6	361
H ₂ Se...H ₂ O	D _{Se}	10.8	9.0	10.8	4.7	5.0	-18.4	125
HCl...H ₂ O	A _H	22.4	19.8	20.9	12.6	15.7	-12.4	220
HClO...H ₂ O	A	30.8	28.4	31.9	23.1	24.6	-3.5	279
HClO ₂ ...H ₂ O	AD _O E _{ClO}	43.0	44.2	51.3	40.6	43.9	4.5	313
HClO ₃ ...H ₂ O	AD _O W _{ClO}	38.7	40.7	49.1	38.2	41.6	2.1	303
HClO ₄ ...H ₂ O	AD _O ² W _{ClO}	47.2	46.9	55.8	45.9	48.4	11.4	338
HClO ₄ ...H ₂ O	AD _O W _{ClO}	45.1	44.6	53.2	43.9	46.2	9.6	285
HBr...H ₂ O	A _H	20.8	15.7	19.1	11.6	14.2	-12.8	198
HBrO...H ₂ O	A	30.3	27.1	31.4	23.0	24.2	-1.6	290
HBrO ₂ ...H ₂ O	AD _O W _{BrO}	46.3	45.2	53.9	43.7	46.8	7.4	315
HBrO ₃ ...H ₂ O	AD _O W _{BrO}	46.6	44.6	54.7	43.8	47.3	7.7	306
HBrO ₄ ...H ₂ O	AD _O W _{BrO}	51.7	49.8	59.3	48.9	52.0	12.7	358
HBrO ₄ ...H ₂ O	AD _O W _{BrO}	50.4	48.3	58.0	48.3	51.2	12.8	360
HI...H ₂ O	A _H	10.2	8.2	(9.3)	(3.6)	(4.9)	(-17.5)	166
HI...H ₂ O	D _I	4.1	(-)	(-)	(-)	(-)	(-)	(-)
HIO...H ₂ O	E _{IO}	22.1	16.6	(19.9)	(14.1)	(14.1)	(-16.7)	241
HIO...H ₂ O	A	25.1	22.2	(27.2)	(19.1)	(20.1)	(-7.0)	101
HIO ₂ ...H ₂ O	AD _O W _{IO}	44.1	45.6	(50.8)	(40.6)	(43.8)	(3.9)	356
HIO ₃ ...H ₂ O	AD _O W _{IO}	49.1	47.7	(55.4)	(43.6)	(48.0)	(5.0)	328
HIO ₄ ...H ₂ O	AD _O W _{IO}	52.8	51.3	(59.7)	(49.5)	(53.0)	(14.8)	388

^a ΔE_e , ΔE_0 , ΔH_r , and ΔG_r in kJ/mol. The temperature (T_h) [K] to form a stable monohydration structure [i.e., $\Delta G(T_h) = 0$] is approximately estimated based on the harmonic frequency thermal energies. The CCSD(T)/CBS values in parentheses are approximate because only the CRENL ECP basis set was employed for iodine without basis set expansion. In the case of monohydrated HI (D_I), it was not possible to converge CCSD(T) calculations.

AD_OW_{SO} is ~2.8 kJ/mol higher. The structures with a H-bond formed between the proton donor in the water molecule and the oxygen atom belonging to the —S=O group (AD_O) are slightly more stable than the AD_{OH} structures. The q_S of H₂SO₃ (2.38 au) is slightly larger than the q_S of (HO)₂SO (1.91 au). The r_{SO} distance in monohydrated H₂SO₃ (3.34 Å) is relatively longer than that in (HO)₂SO (3.08 Å).

In the case of monohydrated sulfuric acid (H₂SO₄), the most stable structure is AD_OW_{SO}, which is slightly more stable than A²D_OW_{SO}. During geometry optimization, the hypothetical AD_O² motif is converted to the most stable AD_OW_{SO} isomer. The most stable structures of monohydrated H₂SO₃ and H₂SO₄

are very close in interaction energy because both conformers are represented by the AD_OW_{SO} motif, as noted in the case of monohydrated H₃PO₃ and H₃PO₄, which are nearly isoenergetic due to the same AD_OW_{PO} motif.

Hydrogen selenide (H₂Se) is structurally similar to the hydrogen sulfide. It forms two different stable conformers with a water molecule; the structure D_{Se} is slightly more stable than the structure A_H by ~1.7 kJ/mol, in contrast to the case of monohydrated hydrogen sulfide (H₂S) for which both structural isomers are very close in energy.

Monohydrated Inorganic Oxyacid Complexes Containing a Halogen Atom. The acids having chlorine, bromine, and

iodine are significantly different from the previous cases. These heavy atoms are highly electronegative and highly polarizable.

In the case of monohydrated hypochlorous acid (HClO), the most stable conformer is A, while E_{ClO} is significantly less stable by ~ 15.9 kJ/mol. The E_{ClO} structure has no hydrogen bond, but it is formed because of relatively strong electrostatic attraction between the negatively charged oxygen atom of water molecule and the positively charged chlorine atom ($q_{\text{Cl}} = 0.16$ au) of hypochlorous acid. In this case, the interatomic distance between Cl and O (2.77 Å) is much shorter in comparison with the sum of the van der Waals radii (3.27 Å), since the electrostatic interaction overwhelms the van der Waals interaction.

For the chlorous acid (HClO₂), the $AD_{\text{O}}E_{\text{ClO}}$ structure is the most stable. For the monohydrated chloric acid (HClO₃), the $AD_{\text{O}}W_{\text{ClO}}$ structure is the most stable. In the case of monohydrated perchloric acid (HClO₄), the $AD_{\text{O}}^2W_{\text{ClO}}$ structure is slightly more stable than the $AD_{\text{O}}W_{\text{ClO}}$ structure by ~ 1.8 kJ/mol. The D_{O} conformer is significantly less stable in comparison with $AD_{\text{O}}^2W_{\text{ClO}}$ and $AD_{\text{O}}W_{\text{ClO}}$ clusters.

The monohydrated hypobromous acid (HBrO) forms conformers similar to the monohydrated hypochlorous acid. The A conformer is the most stable. The E_{BrO} conformer which has no hydrogen bond is less stable by ~ 8.1 kJ/mol. The electrostatic interaction between bromine and oxygen is stronger in comparison with the E_{ClO} structure of monohydrated HClO because the q_{Br} of HBrO is 0.25 au. The monohydrated bromous acid (HBrO₂) forms four different molecular complexes. The most stable conformer is represented by the $AD_{\text{O}}W_{\text{BrO}}$ motif. The D-type structures are significantly less stable in comparison with $AD_{\text{O}}W_{\text{BrO}}$. The bromic acid (HBrO₃) forms a few different structures by interacting with a water molecule; the AD_{O} and D types. The most stable conformer is $AD_{\text{O}}W_{\text{BrO}}$. The D conformers are significantly less stable. We also tried to find A, D_{O} , and D_{OH} motifs. During geometry optimization, all these conformers are converted to different structures (A to $AD_{\text{O}}W_{\text{BrO}}$ and D_{O} as well as D_{OH} are transformed to $D_{\text{OH}}D_{\text{O}}E_{\text{BrO}}$). The monohydrated perbromic acid (HBrO₄) forms two stable $AD_{\text{O}}W_{\text{BrO}}$ conformers. Both clusters are very close in energy. During geometry optimization, hypothetical A, D_{O} , and D_{OH} are converted to the most stable $AD_{\text{O}}W_{\text{BrO}}$ form.

In the case of monohydrated hypoiodous acid (HIO), it is surprising that the most stable structure (E_{IO}) has no hydrogen bond. The E_{IO} conformer is governed by electrostatic interactions. The conformer A is slightly less stable than E_{IO} by ~ 0.6 kJ/mol in ΔE_e and ~ 2.9 kJ/mol in ΔE_0 . In the series of HXO acids where X is Cl, Br, or I, the interaction energy difference between A and E_{XO} structures decreases with the increasing atomic mass of the halogen atom. In the case of hypoiodous acid, this order is even reversed. The relative interaction energy difference in $-\Delta E_e(-\Delta E_0)$ between A and E_{XO} for HClO, HBrO, and HIO is $\sim 20.0(15.9)$, $\sim 11.0(8.1)$, and $\sim -0.6(-2.9)$ kJ/mol, respectively. For the monohydrated iodous acid (HIO₂), the most stable conformer is $AD_{\text{O}}W_{\text{IO}}$. In this case, the H-bond length ($X-\text{OH}\cdots\text{OH}_2$) is slightly shorter than that in HBrO₂. In the case of monohydrated iodic acid (HIO₃) the $AD_{\text{O}}W_{\text{IO}}$ structure is similar to those formed by HBrO₃. We also tried to find A, D_{O} , and D_{OH} conformers for the monohydrated HIO₃. During geometry optimization, A as well as D_{O} are transformed to the most stable $AD_{\text{O}}W_{\text{IO}}$ form, and the D_{OH} structure is converted to a less stable $D_{\text{OH}}D_{\text{O}}E_{\text{IO}}$ isomer. For the periodic acid (HIO₄), the $AD_{\text{O}}W_{\text{IO}}$ structure is the most stable. The $D_{\text{OH}}^2E_{\text{IO}}$ conformer is significantly less stable, however, the

structure of $D_{\text{OH}}^2E_{\text{IO}}$ is determined by strong electrostatic interaction between I and O. During geometry optimization, hypothetical A, D_{O} , and D_{OH} are converted to the most stable $AD_{\text{O}}W_{\text{IO}}$ form.

To verify the reliability of CRENBLECP basis set for iodine, we carried out calculations for $\text{HBrO}_{n=0-4}\cdots\text{H}_2\text{O}$ complexes employing the CRENBLECP basis set for the bromine atom and the aVDZ basis set for oxygen and hydrogen. The calculated values are in good agreement with the results obtained for the same complexes where all the atoms were treated with the aVDZ basis set. In general, the order in stability of conformers is preserved. However, in some cases when the energy differences between conformers are small, the order is changed. The largest binding energy difference (in terms of $-\Delta E_0$) can be observed for monohydrated HBrO₂ ($D_{\text{O}}W_{\text{BrO}}$ motif). In this case, the energy difference is 6.9 kJ/mol. In other cases the corresponding energy differences are substantially smaller. The order is reversed for the $D_{\text{O}}D_{\text{OH}}E_{\text{BrO}}$ and $D_{\text{O}}^2E_{\text{BrO}}$ conformers of monohydrated HBrO₃. However, both structures are almost isoenergetic. For comparison, the table listing the results of $\text{HBrO}_{n=0-4}\cdots\text{H}_2\text{O}$ complexes based on the above basis sets is in the Supporting Information (Table S3).

All reported NBO charges are localized on the acid (Table 3). The NBO charges localized on hydrogen atom ($q_{\text{H}(\text{O})}$) belonging to the $-\text{OH}$ group are similar; the relative differences do not exceed 0.05 au. The NBO charges localized on oxygen atom ($q_{\text{O}(\text{H})}$) are significantly different depending on different chemical species. For metaboric and orthoboric acids, $q_{\text{O}(\text{H})}$ charges are -1.04 and -1.06 au, respectively. Those values are significantly more negative in comparison with nitrous and nitric acids (-0.64 and -0.58 au, respectively). In the case of acids containing phosphorus, the $q_{\text{O}(\text{H})}$ charges are slightly more negative in comparison with the boric acids. The acids containing halogen atoms enable more in-depth analysis of the charge distribution. As the atomic mass of halogen atom increases, the corresponding $q_{\text{O}(\text{H})}$ NBO charge becomes more negative. This is directly related to the NBO charge q_{X} of the halogen atom. As the atomic mass of X increases, the NBO charge localized on the central atom is more positive, thus the X–O bond becomes highly polarized. The q_{X} 's are positive, except for acids which do not contain oxygen atoms. In the case of diprotic acids H_2X (X = S, Se), the q_{X} is more negative for the hydrogen sulfide. The sulfur is more electronegative than selenium, thus the S–H bond is more polarized. In both cases the most stable structures are D_{X} , thus the interaction energy is determined by a proton donor of the water molecule. In the case of monohydrated hydrogen halide acids, the most stable structure is A, since the charge distribution of HX highly affects the interaction energy.

Interaction Energies, Bond Distances, NBO Charges, and IR Red-Shifts of –OH Vibrational Modes with Respect to the Number of Oxygen Atoms (n) in $\text{HXO}_{n=0-4}$ (X = Cl, Br, I). The plots in Figure 2 are based on the minimum-energy structures [$(\text{HBO}_2/\text{A}-\text{H}_3\text{BO}_3/\text{AD}_{\text{OH}})$, $(\text{HNO}_2/\text{A}(\text{D}_{\text{O}})-\text{HNO}_3/\text{AD}_{\text{O}})$, $(\text{H}_3\text{PO}_3-\text{AD}_{\text{O}}/\text{AD}_{\text{O}}\text{W}_{\text{PO}})$, $(\text{H}_2\text{SO}_3-\text{AD}_{\text{O}}/\text{AD}_{\text{O}}\text{W}_{\text{SO}})$, $(\text{HClO}/\text{A}-\text{HClO}_2-\text{AD}_{\text{O}}/\text{AD}_{\text{O}}\text{W}_{\text{ClO}})$, $(\text{HClO}_3-\text{AD}_{\text{O}}/\text{AD}_{\text{O}}\text{W}_{\text{ClO}})$, $(\text{HBrO}/\text{A}-\text{HBrO}_2-\text{AD}_{\text{O}}/\text{AD}_{\text{O}}\text{W}_{\text{BrO}})$, and $(\text{HIO}/\text{A}-\text{HIO}_2-\text{AD}_{\text{O}}/\text{AD}_{\text{O}}\text{W}_{\text{IO}})$]. Figure 2a shows the relationship between binding energy [kJ/mol] and number of oxygen atoms (n) in monohydrated inorganic acids. For monohydrated $\text{HXO}_{n=0-4}$ (X = Cl, Br, I) the binding energy (the negative value of interaction energy) increases up to $n = 2$ and becomes almost constant for $n = 2-4$. In the case of $n = 3$, the binding energy is slightly smaller than those of $n = 2, 4$.

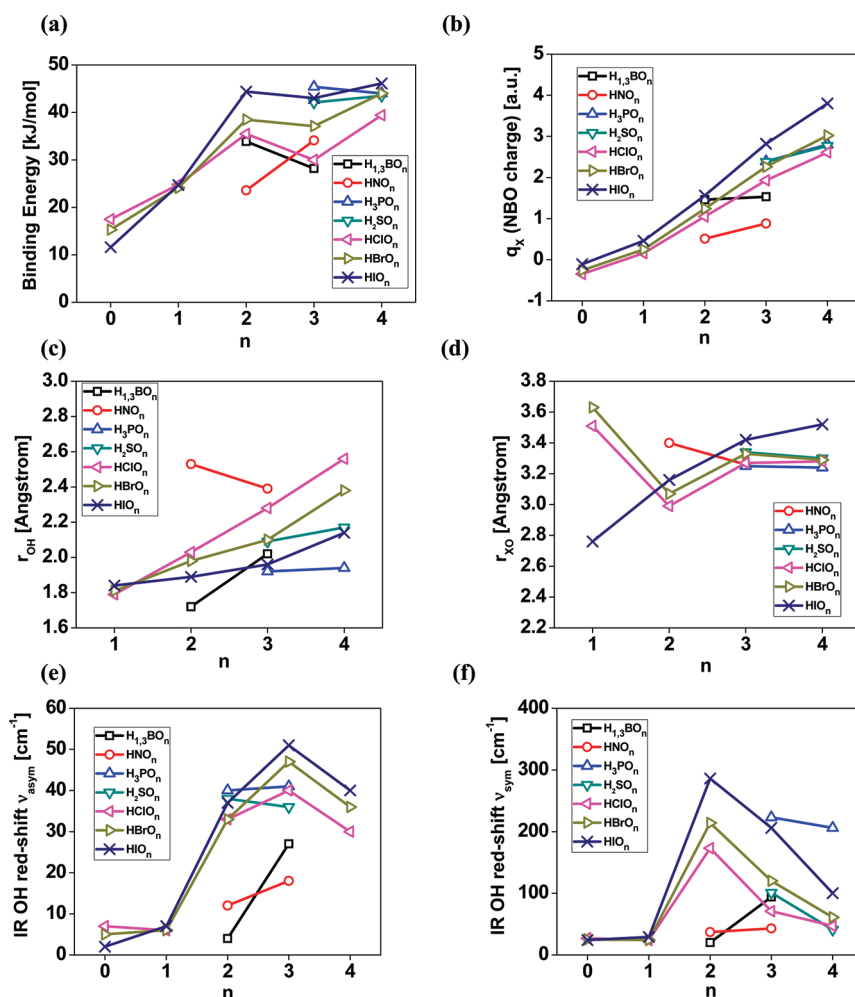


Figure 2. Plots of interaction energy [ΔE_0 in kJ/mol] (a), NBO charge [q_X in au] (b), MP2/aVDZ r_{OH} in Å (c), MP2/aVDZ r_{XO} in Å (d), IR red-shifts of –OH asymmetric (e) and symmetric (f) vibrational frequencies in cm^{-1} , with respect to the number of oxygen atoms (n). $H_{1,3}BO_n$ denotes $HBO_{n=2}$ and $H_3BO_{n=3}$. All plots are based on the minimum-energy structures [($HBO_2/A-H_3BO_3/AD_{OH}$), ($HNO_2/A(D_O)$, HNO_3/AD_O), ($H_3PO_{3-4}/AD_{O}W_{PO}$), ($H_2SO_{3-4}/AD_{O}W_{SO}$), ($HClO/A-HClO_{2-3}/AD_{O}W_{ClO-HClO_4}/AD_{O}^2W_{ClO}$), ($HBrO/A-HBrO_{2-4}/AD_{O}W_{BrO}$), and ($HIO/A-HIO_{2-4}/AD_{O}W_{IO}$)]. The binding energy tends to increase up to $n = 2$ but becomes almost similar for $n = 2-4$; $\nu_3(OH)$ tends to be maximized around $n = 3$; $\nu_1(OH)$ tends to be maximized around $n = 2$. On the other hand, q_X , r_{OH} , and r_{XO} tend to increase with n for the conformations having the same type of motif. When the conformation motif changes significantly, such trends can be changed, as can be noted for the changes in the cases of ($HBO_2-H_3BO_3$), (HNO_{2-3}), ($HClO_{1-2}$), and ($HBrO_{1-2}$).

Figure 2b shows the relationship between the q_X NBO charges and number of oxygen atoms (n) in various inorganic acids. The graph clearly demonstrates the direct proportional relation between q_X and n (except boron containing acids). The NBO charges localized on iodine are relatively more positive than those on bromine and chlorine. The q_N charges localized on HNO_2 and HNO_3 are less positive in comparison with other inorganic acids.

Figure 2c and d shows the r_{OH} and r_{XO} distances. The relationship between r_{OH} and n and that between r_{XO} and n are similar. In both cases, for most systems, the bond length increases with n . On the other hand, the r_{OH} and r_{XO} tend to increase with n for the conformations having the same type of motif. However, when the conformation motif changes significantly, such trends can be changed, as can be noted for the changes in the cases of (HNO_{2-3}), ($HClO_{1-2}$), and ($HBrO_{1-2}$).

Figure 2e and f shows the relationship between the IR red-shifts of –OH stretching vibrational modes (asymmetric and symmetric) of the water molecule and the number of oxygen

atoms (n) for the most stable monohydrated inorganic acids. The asymmetric and symmetric vibrational modes refer to the water molecule. The IR red-shifts for –OH symmetric vibrational modes are much larger than those for –OH asymmetric vibrations, as in water clusters.⁵⁸ The water symmetric OH stretch frequency [$\nu_3(OH)$] tends to be maximized around $n = 3$; the water asymmetric OH stretch frequency [$\nu_1(OH)$] tends to be maximized around $n = 2$. This trend is not valid in the case of ($HBO_2-H_3BO_3$), because the conformation motif is completely different.

For the most stable conformers, we analyze the OH stretching harmonic frequencies calculated at the MP2/aVDZ level of theory, the vibrational analysis is provided to facilitate future experiments. Figure 3 (and Tables S4 and S5 in the Supporting Information) shows the OH stretching frequencies involved in the H-bonding between the acid and the water molecule. The –OH stretching vibrational modes are important in structure identification during infrared photodissociation spectroscopy

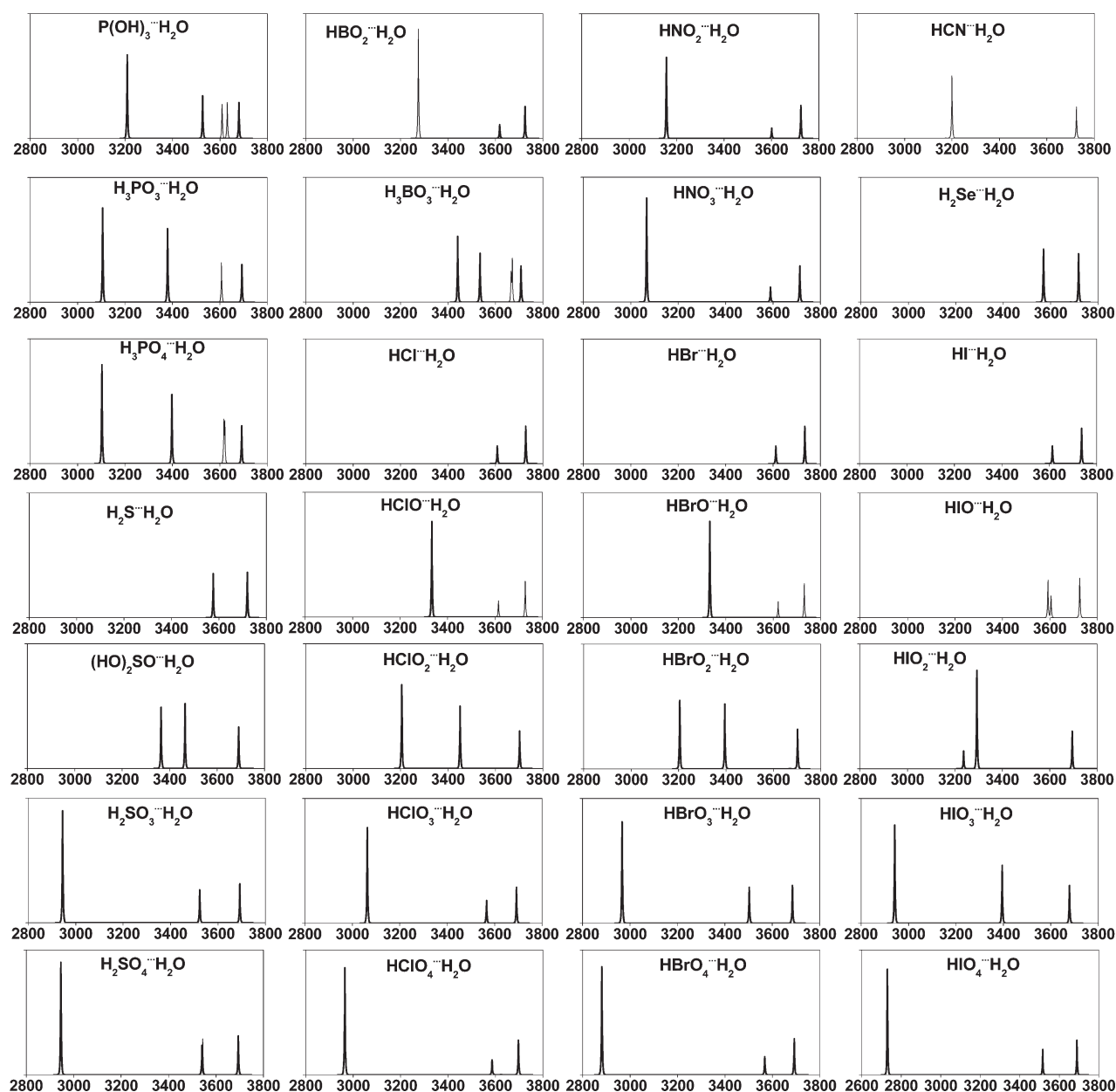


Figure 3. Predicted IR spectra of the OH anharmonic stretching frequencies (MP2/aVDZ values [cm^{-1}]). Thin lines represent stretching $-\text{OH}$ vibrational modes of OH group which is involved in H-bond formation.

(IRPD) experiments. The IR vibrational frequencies show significant differences between acids with and without oxygen atoms. It is interesting to note the higher red-shifts of the $-\text{OH}$ frequencies with increasing number of n for $\text{HXO}_n \cdot \text{H}_2\text{O}$, where $X = \text{Cl}, \text{Br}, \text{I}$. Some IR spectra exhibit coupling modes, which correspond to the simultaneous vibrations of $-\text{OH}$ groups in inorganic acids and water molecules. The analysis of IR spectra of HClO , HBrO , and HIO clearly shows that the minimum-energy structure for HIO significantly differs from those of HClO and HBrO .

The anharmonicity of $-\text{OH}$ stretching modes is crucial for monohydrated inorganic acids. Ab initio calculations carried out with an appropriate wave function model using the harmonic approximation can reasonably reproduce the shift of the $\text{A}-\text{H}$ stretching band upon H-bonding. It is particularly so, if the equilibrium structure is located in a relatively deep potential well,

so that both the $v = 0$ and the $v = 1$ vibrational states of the proton-stretching mode are confined within this well. However, if the equilibrium structure is found in a region of the potential energy surface, which is broad and relatively flat, or if a second region of the surface can be accessed in either the $v = 0$ or the $v = 1$ vibrational state of the proton-stretching mode, then the harmonic approximation fails. The potential energy surfaces for many monohydrated inorganic acids are relatively broad and flat. The analysis of very intensive $-\text{OH}$ stretching modes of inorganic acids clearly shows that the differences between anharmonic and scaled harmonic frequencies (the scaling factor is 0.957)^{4b} for this stretching mode in some cases are $\sim 100 \text{ cm}^{-1}$. Thus, for the most stable conformers, we calculated anharmonic frequencies based on the quartic potential surface using the Gaussian 03 suite of programs. Figure 3 contains the anharmonic

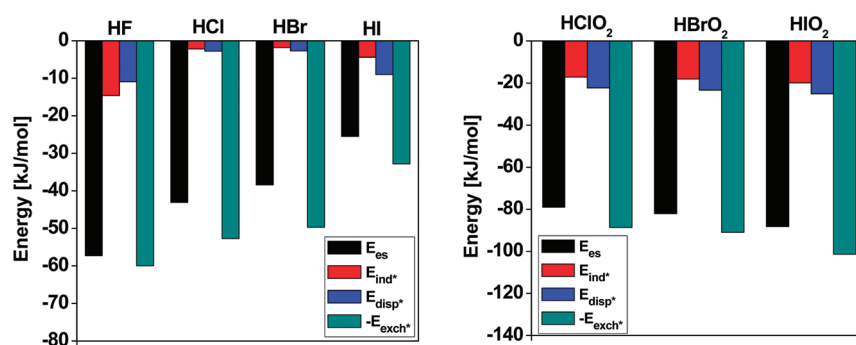


Figure 4. SAPT-DFT energy contributions [kJ/mol] for monohydrated HX ($X = \text{F, Cl, Br, I}$) and monohydrated HXO_2 ($X = \text{Cl, Br, I}$) complexes. E_{es} , E_{ind^*} , E_{disp^*} , and E_{exch^*} denote electrostatic, effective induction, effective dispersion, and effective exchange–repulsion energies, respectively.

frequencies for symmetric and asymmetric $-\text{OH}$ stretching modes. It is important to note that scaled harmonic frequencies are usually in good agreement with anharmonic frequencies.

It is interesting to compare monohydrated $\text{HXO}_{n=2}$ complexes (as a representative case of $\text{HXO}_{n=1-3}$ for convenience's sake) with monohydrated HX complexes ($X = \text{F, Cl, Br, I}$). The SAPT-DFT calculations show that the electrostatic component in the monohydration energy is dominant (E_{es}) due to the H-bonding nature of both complexes (Figure 4). In the case of monohydrated HX clusters, the magnitude of electrostatic energy ($|E_{\text{es}}|$) decreases with increasing size of the halide and so is the exchange energy (E_{exch^*}) because the equilibrium distance is determined mainly by the sum of ($E_{\text{es}} + E_{\text{exch}^*}$) as long as the absolute values of induction ($|E_{\text{ind}^*}|$) and dispersion energies ($|E_{\text{disp}^*}|$) are small. In the case of F, the $\text{F}\cdots\text{H}-\text{O}$ distance is so short that the $|E_{\text{disp}^*}|$ is large, but the E_{exch^*} is even larger. Thus, excluding the case of F, the $|E_{\text{disp}^*}|$ increases with increasing size of the halide, but the value is not significant. On the other hand, in the case of monohydrated HXO_2 clusters the magnitude of electrostatic energy ($|E_{\text{es}}|$) increases with increasing size of the halide in contrast to the case of monohydrated HX clusters, and the dispersion and induction terms are significant. The B3LYP functional does not take into account the dispersion energy, thus in the presence of heavy X atoms, the interaction energies are smaller in magnitude than the MP2 values, and the interatomic distances are relatively longer.

In some cases, the minimum energy conformers are not stable at standard conditions. The stability of these complexes is analyzed on the basis of Gibbs free energy values (ΔG_{T}). In Table 4 we report T_{h} values for the most stable conformers. T_{h} corresponds to temperature at which Gibbs free energy is zero ($\Delta G(T_{\text{h}}) = 0$). The T_{h} values give the information at what temperature a specific acid complex will have the same concentration, as the inorganic acid is no longer bound to water.

CONCLUDING REMARKS

We conducted a comprehensive analysis of different monohydrated inorganic acids. We carried out geometry optimization along with harmonic frequency calculations at the B3LYP and MP2 levels of theory. For each molecular system we studied all possible conformers. For inorganic acids without oxygen atoms, one or two stable conformers are found. In the case of monoprotic hydrogen halide acids, HX ($X = \text{F, Cl}$) one stable structure is found. For heavier halogen atoms, two stable structures are formed. The second structure does not contain H-bonds, the positively charged X atom interacts with the negatively charged

oxygen atom of water molecule. The most stable monohydrated hydrogen halide is formed by the hydrogen fluoride. The monohydrated hydrogen iodide has the smallest binding energy (the negative value of interaction energy). Among monohydrated hydrogen halides, the diprotic acids H_2X ($X = \text{S, Se}$) form two different stable conformers. The $\text{H}_2\text{S}\cdots\text{H}_2\text{O}$ complex is slightly more stable than $\text{H}_2\text{Se}\cdots\text{H}_2\text{O}$. The metaboric acid as well as the orthoboric acid forms one stable structure, while interacting with a single water molecule. In the case of other acids, the number of possible conformers is substantially larger. In the majority of cases, AD structures are preferred. The AD structure has two H-bonds, while both acid and water are proton donors and acceptors. The inorganic acids, including chlorine, bromine, and iodine atoms, form slightly different structures. The most stable conformer in the case of $\text{HIO}\cdots\text{H}_2\text{O}$ complex has no H-bond. The structure is determined by the electrostatic interactions between iodine and oxygen. In many cases, the structures of $\text{HXO}_{n=1-4}$ ($X = \text{Cl, Br, and I}$) are additionally stabilized by noncovalent $\text{X}\cdots\text{O}$ van der Waals interactions, including the $\text{X}^+\cdots\text{O}^-$ electrostatic interactions. The hydronium cation is not formed in the case of monohydrated inorganic acids. In general, the hydronium cation which is formed by proton transfer from the acid to the water molecule is stabilized by other water molecules in the molecular system. For the most stable conformers, we calculated the interaction energy at the CCSD(T)/CBS level of theory. As compared with the CCSD(T)/CBS results, the B3LYP method highly underestimates the binding energies. The MP2/aVDZ values are not significantly different from the CCSD(T)/CBS results; the relative interaction energy differences in the present systems do not exceed ~ 8.2 kJ/mol. The IR spectra of acid \cdots water systems are reported to facilitate the experiments, showing the characteristic features of H-bond interactions.

The detailed analysis of many different monohydrated inorganic acids shows that the complex formation is determined by many different factors. The geometry of the resulting complexes depends on the electronegativity of the central atom, its atomic radius, and the position of the $-\text{OH}$ groups relative to the $\text{X}=\text{O}$ groups. In many cases, not only the noncovalent van der Waals interactions but also the $\text{X}^+\cdots\text{O}^-$ electrostatic interactions additionally stabilize the acid \cdots water complexes.

ASSOCIATED CONTENT

S Supporting Information. Table S1 is the MP2(B3LYP)/aVDZ ZPE-uncorrected/ZPE-corrected interaction energies and thermodynamic properties for various conformers of monohydrated

inorganic acids. Table S2 is the MP2(B3LYP)/aVDZ ZPE-uncorrected/ZPE-corrected binding energies and thermodynamic properties for various conformers of monohydrated inorganic acids containing halogen atoms. Table S3 is the MP2(B3LYP)/aVDZ ZPE-uncorrected/ZPE-corrected interaction energies and thermodynamic properties for monohydrated $\text{HBrO}_{n=0-4}$. Table S4 is the MP2/aVDZ scaled harmonic (scaling factor =0.957) and anharmonic vibrational frequencies for the most stable conformers. Table S5 is the MP2/aVDZ scaled harmonic (scaling factor =0.957) vibrational frequencies for less stable conformers. This material is available free of charges via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kim@postech.ac.kr.

Present Addresses

[†]Department of Theoretical Chemistry, Institute of Chemistry, University of Silesia, 9 Szkolna Street, 40–006 Katowice, Poland.

ACKNOWLEDGMENT

This work was supported by NRF (National Honor Scientist Program, WCU: R32-2008-000-10180-0) and KISTI (KSC-2011-G3-02). We thank referees for the useful comments.

REFERENCES

- (1) (a) Stace, A. *Science* **2001**, *294*, 1292. (b) Miller, D. J.; Lisy, J. M. *J. Chem. Phys.* **2006**, *124*, 184301. (c) Lee, H. M.; Tarakeshwar, P.; Park, J. W.; Kołaski, M. R.; Yoon, Y. J.; Yi, H.-B.; Kim, W. Y.; Kim, K. S. *J. Phys. Chem. A* **2004**, *108*, 2949. (d) Singh, N. J.; Park, M.; Min, S. K.; Suh, S. B.; Kim, K. S. *Angew. Chem., Int. Ed.* **2006**, *45*, 3795. (e) Cooper, T. E.; Carl, D. R.; Armentrout, P. B. *J. Phys. Chem. A* **2009**, *113*, 13727. (f) Carl, D. R.; Chatterjee, B. K.; Armentrout, P. B. *J. Chem. Phys.* **2010**, *132*, 044303.
- (2) (a) Hurley, S. M.; Dermota, T. E.; Hydutsky, D. P.; Castleman, A. W. *Science* **2002**, *298*, 202. (b) Weber, J. M.; Kelley, J. A.; Nielsen, S. B.; Ayotte, P.; Johnson, M. A. *Science* **2000**, *287*, 2461. (c) Kołaski, M.; Lee, H. M.; Pak, C.; Kim, K. S. *J. Am. Chem. Soc.* **2008**, *130*, 103. (d) Majumdar, D.; Kim, J.; Kim, K. S. *J. Chem. Phys.* **2000**, *112*, 101. (e) Xantheas, S. S.; Dunning, T. H. *J. Phys. Chem.* **1994**, *98*, 13489. (f) Yates, B.; Schaefer, H. F., III; Lee, T. J.; Rice, J. E. *J. Am. Chem. Soc.* **1988**, *110*, 6327.
- (3) (a) Gutberlet, A.; Schwaab, G.; Birer, O.; Masia, M.; Kaczmarek, A.; Forbert, H.; Havenith, M.; Marx, D. *Science* **2009**, *324*, 1545. (b) Odde, S.; Mhin, B. J.; Lee, K. H.; Lee, H. M.; Tarakeshwar, P.; Kim, K. S. *J. Phys. Chem. A* **2006**, *110*, 7918–7924. (c) Re, S.; Osamura, Y.; Suzuki, Y.; Schaefer, H. F., III. *J. Chem. Phys.* **1998**, *109*, 973.
- (4) (a) Kumar, A.; Park, M.; Huh, J. Y.; Lee, H. M.; Kim, K. S. *J. Phys. Chem. A* **2006**, *110*, 12484–12493. (b) Lee, H. M.; Tarakeshwar, P.; Kim, K. S. *J. Chem. Phys.* **2004**, *121*, 4657–4664. (c) Lee, H. M.; Kim, D.; Singh, N. J.; Kołaski, M.; Kim, K. S. *J. Chem. Phys.* **2007**, *127*, 164311.
- (5) (a) Godinho, S. S. M. C.; do Couto, P. C.; Cabral, B. J. C. *J. Chem. Phys.* **2005**, *122*, 044316. (b) Odde, S.; Mhin, B. J.; Lee, H. M.; Kim, K. S. *J. Chem. Phys.* **2004**, *121*, 11083. (c) Olleta, A. C.; Lee, H. M.; Kim, K. S. *J. Chem. Phys.* **2006**, *124*, 024321. (d) Singh, N. J.; Yi, H.-B.; Min, S. K.; Park, M.; Kim, K. S. *J. Phys. Chem. B* **2006**, *110*, 3808–3815. (e) Olleta, A. C.; Lee, H. M.; Kim, K. S. *J. Chem. Phys.* **2007**, *126*, 144311.
- (6) (a) Orphal, J.; Kou, Q.; Tchana, F. K.; Pirali, O.; Flaud, J.-M. *J. Mol. Spectrosc.* **2003**, *221*, 239. (b) Rothenberg, G.; Beadnall, R. M. H.; McGrady, J. E.; Clark, J. H. *J. Chem. Soc., Perkin Trans.* **2002**, *2*, 630. (c) Otto, A. H.; Schrader, S.; Steiger, T.; Schneider, M. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 3927.
- (7) Walker, M.; Morrison, C. A.; Allan, D. R. *Phys. Rev. B* **2005**, *72*, 224106.
- (8) D'Auria, R.; Turco, R. P.; Houk, K. N. *J. Phys. Chem. A* **2004**, *108*, 3756.
- (9) Tao, F.-M. *J. Chem. Phys.* **1998**, *108*, 193.
- (10) Nguyen, M.-T.; Jamka, A. J.; Cazar, R. A.; Tao, F.-M. *J. Chem. Phys.* **1997**, *106*, 8710.
- (11) Miller, Y.; Chaban, G. M.; Gerber, R. B. *Chem. Phys.* **2005**, *313*, 213.
- (12) Donaldson, D. J.; Orlando, J. J.; Amann, S.; Tyndall, G. S.; Proos, R. J.; Henry, B. R.; Vaida, V. *J. Phys. Chem. A* **1998**, *102*, 5171.
- (13) Feierabenda, K. J.; Havey, D. K.; Varner, M. E.; Stanton, J. F.; Vaida, V. *J. Chem. Phys.* **2006**, *124*, 124323.
- (14) Zhang, X.; Mereand, E. L.; Castleman, A. W., Jr. *J. Phys. Chem.* **1994**, *98*, 3554.
- (15) Fleming, P. R.; Li, M.-G.; Rizzo, T. R. *J. Chem. Phys.* **1991**, *94*, 2425.
- (16) Sinha, A.; Wal, R. L. V.; Crim, F. F. *J. Chem. Phys.* **1989**, *91*, 2929.
- (17) Sinha, A.; Wal, R. L. V.; Crim, F. F. *J. Chem. Phys.* **1990**, *92*, 401.
- (18) Loerting, T.; Liedl, K. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8874.
- (19) Larson, L. J.; Kuno, M.; Tao, F.-M. *J. Chem. Phys.* **2000**, *112*, 8830.
- (20) Fiacco, D. L.; Hunt, S. W.; Leopold, K. R. *J. Am. Chem. Soc.* **2002**, *124*, 4504.
- (21) Aguzzi, A.; Rossi, M. *J. Phys. Chem. Chem. Phys.* **2001**, *3*, 3707.
- (22) Arstila, H.; Laasonen, K.; Laaksonen, A. *J. Chem. Phys.* **1998**, *108*, 1031.
- (23) Bandy, A. R.; Ianni, J. C. *J. Phys. Chem. A* **1998**, *102*, 6533.
- (24) Beichert, P.; Schrems, O. *J. Phys. Chem. A* **1998**, *102*, 10540.
- (25) Re, S.; Osamura, Y.; Morokuma, K. *J. Phys. Chem. A* **1999**, *103*, 3535.
- (26) Voegelé, A. F.; Tautermann, C. S.; Loerting, T.; Hallbrucker, A.; Mayer, E.; Liedl, K. R. *Chem.—Eur. J.* **2002**, *8*, 5644.
- (27) Bishenden, E.; Donaldson, D. J. *J. Phys. Chem. A* **1998**, *102*, 4638.
- (28) Tarczay, G.; Csaszar, A. G. *J. Chem. Phys.* **2001**, *115*, 1229.
- (29) Pak, C.-H.; Lee, H. M.; Kim, J. C.; Kim, D.-W.; Kim, K. S. *Struct. Chem.* **2005**, *16*, 187.
- (30) Alexeev, Y.; Windus, T. L.; Zhan, C. G.; Dixon, D. A. *Int. J. Quantum Chem.* **2005**, *104*, 379.
- (31) Ruiz-Morales, Y.; Ziegler, T. *J. Phys. Chem. A* **1998**, *102*, 3970.
- (32) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293.
- (33) Corbridge, D. E. C. *Phosphorus: An Outline of its Chemistry, Biochemistry, and Technology*; Elsevier: Amsterdam, The Netherlands, 1995, pp 677–744.
- (34) Holleman, A. F.; Wiberg, E. *Inorg. Chem.*; Academic Press: San Diego, CA, 2001, pp 397–744.
- (35) Malaspina, T.; Fileti, E. E.; Riveros, J. M.; Canuto, S. *J. Phys. Chem. A* **2006**, *110*, 10303.
- (36) Meot-Ner, M.; Speller, C. V. *J. Phys. Chem.* **1989**, *93*, 1697.
- (37) Irvine, W. M.; Morvan, D. B.; Lis, D. C.; Matthews, H. E.; Biver, N.; Criviser, J.; Davies, J. K.; Dent, W. R. F.; Gautier, D.; Godfrey, P. D.; Keene, J.; Lovel, A. J.; Owen, T. C.; Phillips, T. G.; Rauer, H.; Schloer, F. P.; Senay, M.; Young, K. *Nature (London)* **1996**, *383*, 418.
- (38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. A.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.;

Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(39) Lee, S. J.; Chung, H. Y.; Kim, K. S. *Bull. Korean Chem. Soc.* **2004**, *25*, 1061.

(40) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

(41) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208.

(42) Császár, A. G.; Allen, W. D.; Schaefer, H. F., III *J. Chem. Phys.* **1998**, *108*, 9751.

(43) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schutz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO*, a package of ab initio programs, version 2006.1; Institut für Theoretische Chemie, Universität Stuttgart: Stuttgart, Germany, 2006.

(44) Bondi, A. J. *Phys. Chem.* **1964**, *68*, 441.

(45) Jeziorski, B.; Moszyński, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

(46) (a) Jansen, G.; Heßelmann, A. *J. Phys. Chem. A* **2001**, *105*, 646.

(b) Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464. (c) Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319.

(47) (a) Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778.

(b) Heßelmann, A.; Jansen, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5010.

(48) Lee, E. C.; Kim, D.; Jurecka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457.

(49) Heßelmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103–17.

(50) Odde, S.; Mhin, B. J.; Lee, S.; Lee, H. M.; Kim, K. S. *J. Chem. Phys.* **2004**, *120*, 9524.

(51) Kim, J.; Lee, H. M.; Suh, S. B.; Majumdar, D.; Kim, K. S. *J. Chem. Phys.* **2000**, *113*, 5259.

(52) Dyke, T. R.; Muentner, J. S. *J. Chem. Phys.* **1974**, *60*, 2929.

(53) Dyke, T. R.; J. S. Muentner, J. S. *J. Chem. Phys.* **1972**, *57*, 5011.

(54) Dyke, T. R.; Mack, K. M.; Muentner, J. S. *J. Chem. Phys.* **1977**, *66*, 498.

(55) Odutola, J. A.; Dyke, T. R. *J. Chem. Phys.* **1980**, *72*, 5062.

(56) Kim, K. S.; Mhin, B. J.; Choi, U.-S.; Lee, K. *J. Chem. Phys.* **1992**, *97*, 6649.

(57) Keutsch, F. N.; Goldman, N.; Harker, H. A.; Leforestier, C.; Saykally, R. J. *Mol. Phys.* **2003**, *101*, 3477.

(58) Lee, H. M.; Suh, S. B.; Lee, J. Y.; Tarakeshwar, P.; Kim, K. S. *J. Chem. Phys.* **2000**, *112*, 9759.