JCTC Journal of Chemical Theory and Computation

Efficient Diffuse Basis Sets: cc-pV*x*Z+ and maug-cc-pV*x*Z

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

Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431

Received December 28, 2008

Abstract: We combine the diffuse basis functions from the 6-31+G basis set of Pople and co-workers with the correlation-consistent basis sets of Dunning and co-workers. In both wave function and density functional calculations, the resulting basis sets reduce the basis set superposition error almost as much as the augmented correlation-consistent basis sets, although they are much smaller. In addition, in density functional calculations the new basis sets, called cc-pV*x*Z+ where x = D, T, Q, ..., or x = D+d, T+d, Q+d, ..., give very similar energetic predictions to the much larger aug-cc-pV*x*Z basis sets. However, energetics calculated from correlated wave function calculations are more slowly convergent with respect to the addition of diffuse functions. We also examined basis sets with the same number and type of functions as the cc-pV*x*Z+ sets but using the diffuse exponents of the aug-cc-pV*x*Z basis sets and found very similar performance to cc-pV*x*Z+; these basis sets are called minimally augmented cc-pV*x*Z, which we abbreviate as maug-cc-pV*x*Z.

1. Introduction

The choice of basis set for electronic structure calculations is a key factor in determining both the reliability of a calculation and its cost.[1] In recent years, the correlation consistent (cc) basis sets of Dunning and co-workers[2−6] have become very popular. These are arranged in two series, in particular cc-pV*x*Z and aug-cc-pV*x*Z with *x*Z = DZ ("double-ζ"), TZ ("triple-ζ"), QZ ("quadruple-ζ"), etc. The prefix aug- (for "augmented") denotes the addition of diffuse basis functions, i.e., Gaussian functions with small exponential parameters times spherical harmonics. In the aug-cc-pV*x*Z basis sets, one adds quite a few diffuse functions, e.g., in aug-cc-pVDZ one adds 4 diffuse functions on each H and 9 on each heavy (i.e., non-hydrogenic) atom, whereas in aug-cc-pVTZ one adds 9 diffuse functions on each H and 16 on each heavy atom. Another popular set of basis functions is 6-31+G(d)[7] and 6-311+G(2df,p) of Pople and co-workers[8] in which the single plus sign denotes the addition of no diffuse functions on hydrogens and 4 on each heavy atom. (The same Gaussian exponential parameters are used for the diffuse basis functions in these two bases.) We have recently had some success combining the plus strategy with the cc basis sets, yielding cc-pVDZ+, cc-pVTZ+, and cc-pVQZ+, and so we decided to make some systematic comparisons of the performance of these strategies when applied to the same test problems. Those systematic comparisons are the primary subject of this article. We also show that for density functional calculations very similar results can be obtained by drastically pruning the aug-cc-pV*x*Z basis sets; this latter strategy is called minimal aug-cc-pV*x*Z, abbreviated maug-cc-pV*x*Z.

Table 1. Number of Contracted Basis Functions for $H_2SO_4(NH_3)$

	x = D	x = D+d	x = T	x = T+d	x = Q	x = Q+d
cc-pV*x*Z	113	118	254	259	484	489
cc-pV*x*Z+	137	142	278	283	508	513
maug-cc-pV*x*Z	137	142	278	283	508	513
aug-cc-pV*x*Z	187	192	395	400	714	719

The above considerations also extend to a sequence[6] of basis sets with *x*Z = (D+d)Z, (T+d)Z, etc. in which five extra *d* functions are added to each element from Al to Ar. Table 1 compares the number of basis functions in the various series of basis sets for one of our test cases, namely the complex of H_2SO_4 with NH_3 (similar trends would be observed for other sample molecules). The table shows that tremendous cost savings are attainable with the plus basis sets, if they are accurate enough. Are they?

2. Tests

Several issues were considered in choosing and designing our tests. First, diffuse functions are more important, in general, for energetics than for geometries and are more important for electron affinities, barrier heights, and noncovalent interactions than for ionization potentials and bond energies. Second, when simply examining the absolute accuracy of some computed properties, such as geometries or strengths of noncovalent interactions, the improvement in the basis set due to adding diffuse functions may be masked by other aspects of the calculations that lead to larger errors than the basis set's incompleteness. Third, a special aspect of basis set incompleteness that demands attention, especially in the computation of

* Corresponding author e-mail: truhlar@umn.edu.

10.1021/ct800575z CCC: $40.75
Published on Web 03/31/2009

noncovalent interactions, is basis set superposition error (BSSE),[9,10] and if this is not small it is not even clear *how* to extract an estimate of the interaction energy from the calculations. For example, should one include a counterpoise correction (*CpC*) and how should that correction be defined? The *CpC* becomes not only ambiguous but also costly when one considers clusters containing more than two molecules,[11] and we therefore wish to keep it small and avoid it.

With these considerations in mind, we decided to carry out six sets of test calculations. In the first four we compare calculated ionization potentials, electron affinities, bond energies, and barrier heights to databases of accurate (or at least best available) estimates. In the fifth and sixth we examine the size of the basis set superposition error for a set of noncovalent complexes.

The databases used for the first four tests are as follows:

• ionization potentials: database IP13/3,[12,13] containing C, S, O, Si, P, Cl, OH, PH, SH, PH_2, O_2, S_2, and Cl_2;

• electron affinities: database EA13/3,[12,13] containing the same 13 species;

• atomization energies: database AE6,[14] containing SiH_4, SiO, S_2, propyne, glyoxal, and cyclobutane;

• barrier heights: database DBH24/08,[15] consisting of 4 subdatabases: HATBH6, NSBH6, UABH6, and HTBH6 containing 6 barrier heights each (forward and reverse barrier heights of three reactions) for heavy-atom transfer, nucleophilic substitution, unimolecular and association reactions, and hydrogen transfer, respectively.

As explained in previous work, all energetic data in these databases are zero point exclusive, and all calculations for comparison to these databases use standard (QCISD/MG3) geometries.[12−15] All the results presented here include spin−orbit contributions[16] for the C, O, F, Si, S, Cl, OH, and HS species.

For the fifth and sixth test, we studied ten complexes, which are shown in Figure 1. For each complex we calculated the *BSSE* by the functional counterpoise method[9] of Boys and Bernardi and expressed it as a percentage

$$\%CpC = \frac{|CpC|}{BE} \times 100\% \qquad (1)$$

where *BE* is the *CpC*-corrected binding energy. To avoid comparing calculations at different geometries (geometry can have a large effect on BSSE), all *CpC* and *BE* calculations were carried out at standard geometries. In particular, the geometries for $Cl_2(NH_3)$, $HCl(H_2S)$, $(C_2H_4)_2$, and $(CH_4)_2$ were taken from the NCCE31/05 database,[17−19] the geometry of oa-$H_2SO_4(NH_3)$ complex is a nonstationary point selected to show a particular interaction motif, and the others were optimized by the M06-2X[19] density functional with the MG3S[12] basis set. The geometries that are not published elsewhere are given in the Supporting Information.

Because basis set requirements are different for density functional theory[20] (DFT) and wave function theory[21] (WFT), we made calculations with both. For density functionals we consider B3LYP, [22] M05-2X,[18] M06,[19] M06-2X,[19] M06-L,[23] and M06-HF.[24] Note that M06-L is a local density functional, and the others are all nonlocal (also called hybrid). For WFT, we consider Hartree−Fock[25] (HF), second-order perturbation theory[26] (MP2), coupled cluster theory with single and double excitations[27] (CCSD), and CCSD with quasiperturbative connected triple excitations[28] (CCSD(T)).

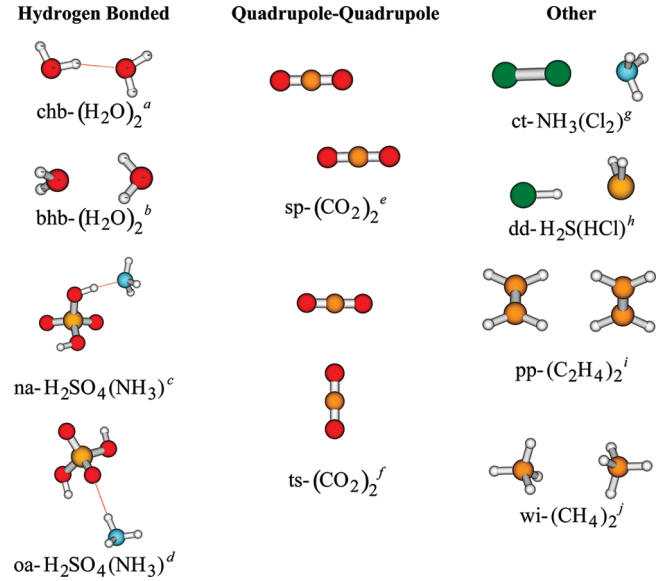


Figure 1. Ten complexes used for the BSSE tests: (a) conventionally hydrogen-bonded (chb) water dimer, (b) bifurcated hydrogen-bonded (bhb) water dimer, (c) hydrogen-bonded dimer of sulfuric acid and ammonia where nitrogen is a hydrogen acceptor (na), (d) hydrogen-bonded dimer of sulfuric acid and ammonia where oxygen is a hydrogen acceptor (oa), (e) slipped parallel (sp) carbon dioxide dimer, (f) T-shaped (ts) carbon dioxide dimer, (g) charge trasfer (ct) dimer of chlorine and ammonia, (h) dipole-dipole (dd) hydrogen chloride and sulfur hydride dimer, (i) π-π stacking (pp) ethene dimer, and (j) weak interaction (wi) methane dimer.

Table 2. Most Diffuse Exponents of Various Symmetries

l	basis set	H	C	N	O	S	Cl
s	cc-pVTZ	0.1027	0.1285	0.1787	0.2384	0.1322	0.1625
	cc-pVTZ+	0.1027	0.0438	0.0639	0.0845	0.0405	0.0483
	aug-cc-pVTZ	0.02526	0.04402	0.0576	0.07376	0.0497	0.0591
	maug-cc-pVTZ	0.1027	0.04402	0.0576	0.07376	0.0497	0.0591
p	cc-pVTZ	0.388	0.1209	0.1725	0.214	0.1098	0.1301
	cc-pVTZ+	0.388	0.0438	0.0638	0.0845	0. 0405	0.0483
	aug-cc-pVTZ	0.102	0.03569	0.0491	0.05974	0.0351	0.0419
	maug-cc-pVTZ	0.388	0.03569	0.0491	0.05974	0.0351	0.0419
d	cc-pVTZ	1.057	0.318	0.469	0.645	0.269	0.344
	cc-pVTZ+	1.057	0.318	0.469	0.645	0.269	0.344
	aug-cc-pVTZ	0.247	0.1	0.151	0.214	0.101	0.135
	maug-cc-pVTZ	1.057	0.318	0.469	0.645	0.269	0.344

All density functional calculations were performed using the *Gaussian 03*[29] program and the MN-GFM[30] functional module. CCSD(T) calculations were carried out by *Molpro-2006.1*.[31]

3. Diffuse Functions

Since the exponential diffuse functions originally developed[7] for use with the 3-21G, 6-21G, and 6-31G basis sets were later found[8] to also be appropriate for use with the 6-311G basis set, we form the new "plus" basis sets by simply adding a diffuse *sp* shell to each non-hydrogenic atom, using the standard exponents that are used in both the 6-31+G and 6-311+G basis sets. To give the reader some idea of how these compare to the most diffuse exponents in the augmented correlation-consistent basis sets, Table 2 compares the exponents of the most diffuse *s*, *p*, and *d* functions in the triple-ζ basis set series under

Table 3. Mean Unsigned Errors (MUEs) (in kcal/mol) in Ionization Potentials

	B3LYP	M06-2X	CCSD(T)
cc-pVDZ	3.14	2.98	10.23
cc-pVDZ+	4.63	2.87	8.85
aug-cc-pVDZ	4.77	2.96	5.68
cc-pVTZ	4.15	2.39	3.44
cc-pVTZ+	4.59	2.63	3.15
aug-cc-pVTZ	4.67	3.45	1.93

Table 4. Mean Unsigned Errors (MUEs) (in kcal/mol) in Electron Affinities

	B3LYP	M06-2X	CCSD(T)
cc-pVDZ	20.84	18.77	30.98
cc-pVDZ+	3.05	4.05	10.49
aug-cc-pVDZ	3.03	3.69	5.27
cc-pVTZ	10.14	8.35	15.05
cc-pVTZ+	2.34	3.06	5.34
aug-cc-pVTZ	2.51	2.98	2.01

consideration here. We see that the exponential parameters of *s* and *p* diffuse functions of the plus and aug basis sets are similar.

The following question arises: what is a "diffuse" basis function? In practice the exponent of the most diffuse valence basis function in the cc-pV*x*Z series decreases as *x* increases. A similar trend is found in even-tempered basis sets.[32] Eventually, one might wonder, does not the most diffuse basis function become indistinguishable from a "diffuse" basis function? In principle, yes, but in practice, no. For example consider the most diffuse *s* on carbon in the cc-pV*x*Z bases. For $x = 2-6$, these are 0.1596, 0.1285, 0.1111, 0.1019, and 0.08635,[33] respectively, whereas the added diffuse *s* function in the plus set has exponent 0.0438, and the added diffuse *s* functions in the aug sets have exponents of 0.0469, 0.04402, 0.04145, 0.0394, and 0.0354. Thus there is a wide gap (0.086 to 0.044) between the most diffuse valence basis function and the added "diffuse" basis function, and one would have to go to unrealistically high *x* before there is any ambiguity; so the definition of "diffuse" is operationally clear.

One could imagine optimizing diffuse functions for various properties such as atomic or molecular Rydberg states, atomic or molecular electron affinities, atomic or molecular polarizabilities, atomic or molecular electron densities, and so forth. One would obtain different results by optimizing against different property databases, but the optimum diffuse functions by any of these approaches would be expected, based on experience, to be a factor of $\sim 2-4$ smaller than the exponent of the least diffuse valence basis function.

4. Results and Discussion

Tables 3−6 give errors relative to the databases[12−15] explained in Section 2.

Tables 3 and 4 show results for ionization potentials and electron affinities. The DFT errors are almost the same for the plus and aug basis sets, even for electron affinities. The coupled cluster calculations, however, are more slowly convergent with respect to the addition of diffuse functions. In order to illustrate this even more clearly, we added the heavy-atom diffuse *d* functions of the aug-cc-pVTZ basis to cc-pVDZ+. The CCSD(T) MUE dropped to only 2.73 kcal/mol for ionization potentials and 3.57 kcal/mol for electron affinities. Although diffuse functions are important for DFT,[12] it is apparently not necessary to exhaust the diffuse space as seems to be required in WFT.

Table 5. Mean Unsigned Errors per Bond (MUEPBs) (in kcal/mol) in Atomization Energies[a]

	B3LYP	M06-2X	CCSD(T)
cc-pVDZ	2.80	2.07	8.87
cc-pVDZ+	3.16	2.38	8.87
aug-cc-pVDZ	2.81	2.13	8.07
cc-pVTZ	0.85	0.40	3.12
cc-pVTZ+	0.94	0.44	3.11
aug-cc-pVTZ	0.94	0.42	2.62

[a] The MUE in atomization energies is divided by the average number (4.83) of bonds per molecule in the AE6 database.

Table 6. Mean Unsigned Errors (MUEs) (in kcal/mol) in the Barrier Heights of the DBH24/08 Database

	HATBH6	NSBH6	UABH6	HTBH6	DBH24
B3LYP/cc-pVDZ	8.63	12.89	2.21	6.48	7.55
B3LYP/cc-pVDZ+	7.45	3.91	2.02	5.98	4.84
B3LYP/aug-cc-pVDZ	7.53	4.15	2.17	5.71	4.89
B3LYP/cc-pVTZ	6.86	9.59	1.65	4.88	5.75
B3LYP/cc-pVTZ+	6.69	4.23	1.71	4.68	4.33
B3LYP/aug-cc-pVTZ	6.81	3.32	1.70	4.71	4.13
M06-2X/cc-pVDZ	2.78	9.64	1.29	1.94	3.91
M06-2X/cc-pVDZ+	1.90	1.34	1.24	1.53	1.50
M06-2X/aug-cc-pVDZ	1.46	1.25	1.35	1.27	1.33
M06-2X/cc-pVTZ	1.08	5.21	1.02	1.43	2.18
M06-2X/cc-pVTZ+	0.77	0.76	1.13	1.30	0.99
M06-2X/aug-cc-pVTZ	0.67	0.66	1.10	1.30	0.93
CCSD(T)/cc-pVDZ	3.94	7.94	1.40	3.03	4.08
CCSD(T)/cc-pVDZ+	4.35	1.08	1.44	1.76	2.16
CCSD(T)/aug-cc-pVDZ	2.27	1.99	0.68	0.76	1.42
CCSD(T)/cc-pVTZ	1.90	5.27	0.47	1.41	2.26
CCSD(T)/cc-pVTZ+	1.65	0.81	0.55	0.99	1.00
CCSD(T)/aug-cc-pVTZ	0.95	0.78	0.33	0.70	0.69

Table 5 shows results for atomization energies. Here the diffuse functions have only a small effect. Table 6, however, shows a large effect for barrier heights, especially for nucleophilic substitution reactions (which is expected since these reactions involve anions). In all DFT cases (all four subdatabasis, two density functionals, and both double and triple-ζ tests), the plus basis sets account quite well for the effect of diffuse functions seen with the larger aug basis sets. They are less successful for the WFT tests, but they often pick up a good fraction of the diffuse function increment at much less cost than is required for aug calculations.

A specific example of barrier height calculations is presented in Table 7. The forward and reverse barrier heights of one of the reactions in the DBH24/08 database were calculated using M06-2X density functional with cc-pV*x*Z, cc-pV*x*Z+, and aug-cc-pV*x*Z (x = T and Q) basis sets. Adding diffuse *s* and *p* functions to the heavy atoms changes the calculated barrier height by 4.64−9.46 kcal/mol. The additional effect of adding diffuse functions on hydrogens and higher angular momentum diffuse functions (*d* and *f* when x = T and *d*, *f*, and *g* when x = Q) on heavy atoms provides a further change in barrier heights of only 0.70−0.85 kcal/mol. The numbers of contracted Gaussian basis functions for this transition state are 146, 158, and 230 for the cc-pVTZ, cc-pVTZ+, and aug-cc-pVTZ basis sets, respectively; these numbers are 285, 297, and 424 for the

Table 7. Barrier Heights (kcal/mol) of the Reaction $OH^- + CH_3F \leftrightarrow HOCH_3 + F^-$ Calculated Using cc-pV*x*Z, cc-pV*x*Z+, and aug-cc-pV*x*Z (*x* = T and Q) Basis Sets with the M06-2X Density Functional[a]

basis set	V_f	V_r
cc-pVTZ	−12.71	8.67
cc-pVTZ+	−3.25	17.52
aug-cc-pVTZ	−2.40	18.28
cc-pVQZ	−8.38	12.91
cc-pVQZ+	−2.37	17.85
aug-cc-pVQZ	−1.60	18.55

[a] V_f is the forward barrier height, and V_r is the reverse barrier height.

Table 8. *%CpC* Values Averaged Over Results from Five Levels of DFT, Four Levels of WFT, and Two Geometries of $H_2SO_4(NH_3)$

basis set	average *%CpC*
cc-pVTZ	17.0
cc-pV(T+d)Z	17.0
cc-pVTZ+	8.1
cc-pV(T+d)Z+	8.2
aug-cc-pVTZ	8.9
aug-cc-pV(T+d)Z	8.7

cc-pVQZ, cc-pVQZ+, and aug-cc-pVQZ basis sets, respectively. The relative computational costs of the cc-pV*x*Z+ basis sets are 1.2 for *x* = T and 1.1 for *x* = Q. In the case of the aug-cc-pV*x*Z basis sets they are 4.4 for *x* = T and 5.5 for *x* = Q. These costs are calculated as the ratio of the cost of a single-point energy calculation for the plus or aug basis set relative to the cost for the corresponding plain basis sets (cc-pV*x*Z); calculations were performed on a Linux cluster with two dual-core 2.6 GHz AMD Opteron processors. Therefore the plus basis sets take account of the main effect of diffuse functions for these barrier height calculations without significantly increasing computational costs.

Next we turn to BSSE. The first BSSE test is restricted to the two $H_2SO_4(NH_3)$ complexes. We averaged the *%CpC* over the two geometries, with five levels of DFT (M05-2X and all four members of the M06 functional family) and four levels of WFT (HF, MP2, CCSD, and CCSD(T)) for a total of 18 cases for each basis set. The results are in Table 8, and they are dramatic; either the plus or the aug basis sets reduce *%CpC* by a factor of 2. With DFT, the computer time for the aug calculations in Table 8 is about 4 times larger than the cc-pVTZ calculations, whereas the time for the cc-pVTZ+ calculations is only 1.5 times larger than that for cc-pVTZ with the four hybrid density functionals and almost the same as cc-pVTZ with M06-L.

The cc-pV(T+d)Z+ basis is very similar in quality to the MG3S basis set[12] that our group has used successfully in many applications. For example, MG3S has 269 contracted functions vs 259 (Table 1) for cc-pV(T+d)Z+, and they have identical diffuse *s* and *p* exponents (Table 2). However MG3S gives 13.5% for the test in Table 8, so cc-pVTZ+ should be considered as a possible improvement for calculating noncovalent interactions.

The final test involves averaging *%CpC* over all ten complexes with the M06-2X density functional and the HF, MP2, CCSD, and CCSD(T) levels of WFT, for a total of 50 cases for each basis set. The results are in Table 9. Here the cc-pV(T+d)Z+ basis set reduces *%CpC* by 36%, whereas aug-cc-pV(T+d)Z reduces it by 49%, averaged over the 50 cases.

Table 9. Average *%CpC* Values Over Four Levels of WFT and One Level of DFT for All Ten Complexes

basis set	average
cc-pV(T+d)Z	39.4
cc-pV(T+d)Z+	25.3
aug-cc-pV(T+d)Z	20.2

Table 10. Mean Unsigned Errors (MUEs) (in kcal/mol) in the Barrier Heights of the DBH24/08 Database Using maug-cc-pVDZ and maug-cc-pVTZ Basis Sets[a]

	HATBH6	NSBH6	UABH6	HTBH6	DBH24
B3LYP/maug-cc-pVDZ	7.66	3.31	2.02	5.83	4.71
M06-2X/maug-cc-pVDZ	1.95	1.25	1.23	1.39	1.46
CCSD(T)/maug-cc-pVDZ	4.15	0.99	1.44	1.85	2.11
B3LYP/maug-cc-pVTZ	6.71	3.28	1.71	4.68	4.09
M06-2X/maug-cc-pVTZ	0.73	0.73	1.14	1.29	0.97
CCSD(T)/maug-cc-pVTZ	1.70	0.30	0.55	1.00	0.89

[a] The maug-cc-pV*x*Z basis set is the aug-cc-pV*x*Z with all diffuse functions removed except for the *s* and *p* diffuse functions on non-hydrogenic atoms.

Table 11. Counterpoise Correction (*CpC*, kcal/mol) from M06-2X Calculations for Various Dimers Using the cc-pV(T+d)Z+ and maug-cc-pV(T+d)Z Basis Sets

dimer	cc-pV(T+d)Z+	maug-cc-pV(T+d)Z[a]	Δ*CpC*[b]
chb-$(H_2O)_2$	0.16	0.12	−0.03
bhb-$(H_2O)_2$	0.25	0.24	−0.02
na-$H_2SO_4(NH_3)$	0.38	0.35	−0.02
oa-$H_2SO_4(NH_3)$	0.11	0.11	0.00
sp-$(CO_2)_2$	0.09	0.06	−0.03
ts-$(CO_2)_2$	0.08	0.06	−0.02
ct-$NH_3(Cl_2)$	0.21	0.23	0.02
dd-$H_2S(HCl)$	0.13	0.10	−0.02
pp-$(C_2H_4)_2$	0.06	0.06	0.00
wi-$(CH_4)_2$	0.03	0.03	0.00
MSD[c]			−0.01

[a] The maug-cc-pV(T+d)Z basis set is the aug-cc-pV(T+d)Z with all diffuse functions removed except for the *s* and *p* diffuse functions on non-hydrogenic atoms. [b] This column gives the energy difference (in kcal/mol) between the two previous columns; the cc-pV(T+d)Z+ *CpC* result is subtracted from the maug-cc-pV(T+d)Z *CpC* one. [c] MSD stands for "mean signed difference" and is the average of the Δ*CpC* values (in kcal/mol).

So far these tests have been applied to the plus basis sets. Next we consider whether similar performance would be obtained with the maug-cc sets. (One can anticipate, especially from the discussion in Section 3, that the performance *will* be similar, but it may be useful to quantitatively validate this.) Table 10 shows some results, similar to Table 6, for maug-cc basis sets; the results for maug-cc-pVDZ are similar to those in Table 6 for the plus basis sets. The results for maug-cc-pVDZ and maug-cc-pVTZ are very similar for reactions in the DBH24/08 database to those for cc-pVDZ+ and cc-pVTZ+ basis sets. For *%CpC* we tested the maug-cc-pV(T+d)Z basis set; a comparison of the cc-pV(T+d)Z+ basis set with the maug-cc-pV(T+d)Z basis set for the binding energies and basis set superposition errors of the dimers shown in Figure 1 is given in Table 11. Again, we see a very similar performance for cc+ and maug-cc basis sets, as expected.

A strategy similar to the plus and maug strategies is to delete all diffuse functions of the aug-cc-pV*x*Z basis sets on hydrogenic atoms and keep all the diffuse functions on non-hydrogenic atoms.[34−36] Another strategy is to delete the highest-angular-momentum diffuse functions on both hydrogenic atoms and non-hydrogenic atoms. Both these strategies have been employed more widely than the plus or maug strategy, although they both lead to much larger basis sets (for a given *x*) than plus or maug. Note that reducing the number of basis functions for a given *x* either decreases computation time for that *x* or allows the use of higher *x* and therefore increases accuracy.

The effect of diffuse functions on *CpC* has also been examined previously.[35−37]

It is beyond the scope of the present study to review the interesting theoretical issues underlying the applicability (or not) of DFT to electron affinities, but the interested reader can consult a review[40] for this. Jensen,[41] in a more recent study, examined the importance of diffuse *s*, *p*, and *d* functions for computing electron affinities, multipole moments, and polarizabilities by DFT and found that diffuse *s* and *p* functions are more important than diffuse *d* functions, but diffuse *d* functions are sometimes important. He also pointed out the difficulty or impossibility of optimizing the diffuse functions in certain cases.

Selective pruning of unaugmented cc-pV*x*Z basis sets has also been examined. Wilson and co-workers examined such truncations[38,39] and showed, for example, that *h* and *g* functions can often be deleted from the cc-pV5Z basis set with only a small loss of accuracy for energies or geometries.

5. Concluding Remarks

While the addition of diffuse functions is necessary, in many cases only diffuse functions of low angular momentum (i.e., diffuse *s* and *p* functions) are needed, and furthermore they are often needed only on heavy atoms. The cc-pV*x*Z+ basis sets are designed to take advantage of this, and as a consequence their basis set size and cost are much smaller than those of the aug-cc-pV*x*Z basis sets, which include diffuse functions of both low and high angular momentum on all atoms. However, the quality of DFT results obtained by using cc-pVTZ+ is much closer to that of aug-cc-pVTZ than to that of cc-pVTZ. Similarly, results for cc-pVDZ+ are closer to those of aug-cc-pVDZ than to those of cc-pVDZ. In fact, in many cases the plus basis sets perform equally well as the aug ones. The benefits of using cc-pV*x*Z+ basis sets with WFT are less dramatic than in the case of DFT. However, they can serve as a good intermediate step between cc-pVTZ and aug-cc-pVTZ, and in fact they are often affordable where the full aug basis is not.

Acknowledgment. The authors are grateful to Steven Mielke for discussions. This work was supported by the United States Department of Energy, Office of Basic Sciences under grant no. DE-FG02-86ER13579.

Supporting Information Available: Five geometries of complexes optimized by M06-2X/MG3S and one [oa-$H_2SO_4(NH_3)$] contrived geometry for the dimers shown in Figure 1 and geometries of the corresponding individual gas-phase monomers with which the binding energies of those dimers were calculated. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Davidson, E. R.; Feller, D. *Chem. Rev.* **1986**, *86*, 681.

(2) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(3) Kendall, R. A.; Dunning, T. H.; Harrision, R. J. *J. Chem. Phys.* **1992**, *96*, 6796. Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(4) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358.

(5) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 2975.

(6) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.

(7) Clark, T.; Chandresekhar, J.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.

(8) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(9) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(10) Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418. Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024. Errata: 1987, *86*, 3760.

(11) Valiron, P.; Mayer, I. *Chem. Phys. Lett.* **1997**, *275*, 46.

(12) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.

(13) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898.

(14) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *107*, 8996. Erratum: *J. Phys. Chem. A* **2004**, *108*, 1460.

(15) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 569. Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808–821.

(16) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643.

(17) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.

(18) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

(19) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(20) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(21) Raghavachari, K.; Anderson, J. B. *J. Phys. Chem.* **1996**, *100*, 12960.

(22) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(23) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1.

(24) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(25) Roothaan, C. C. *J. Rev. Mod. Phys* **1951**, *23*, 69.

(26) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(27) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.

(28) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.

(29) 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.; Zakrze-

wski, 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. *Gaussian03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(30) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module−version 3.0*; University of Minnesota: Minneapolis, MN, 2007.

(31) Werner, H.; Knowles, P. J.; Scheutz, M.; Lindh, R.; Celani, P.; Korona, T.; Rauhut, G.; Manby, F. R.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Ecker, F. H. C.; Hertzer, A.; Lloyd, W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Torroni, R.; Thorsteinsson, T. *1997, MOLPRO 2006.1*; University College Consultants Ltd.: Cardiff, Wales, UK, 2006.

(32) Bardo, R. D.; Ruedenberg, K. *J. Chem. Phys.* **1973**, *59*, 5956.

(33) Wilson, A. K.; van Mourik, T.; Dunning, T. H., Jr. *THEOCHEM* **1997**, *388*, 339.

(34) Sinnokrot, M. O.; Sherill, C. D. *J. Phys. Chem. A* **2003**, *107*, 8377.

(35) Hwang, R.; Muh, S. B.; Lee, J. S. *Mol. Phys.* **2003**, *101*, 1429.

(36) Elsohly, A. M.; Tschumpter, G. S. *Int. J. Quantum Chem.* **2009**, *109*, 91.

(37) Halkier, A.; Koch, H.; Jorgensen, P.; Christiansen, O.; Nielsen, I. M. B.; Helgaker, T. *Theor. Chem. Acc.* **1997**, *97*, 150.

(38) Prascher, B. P.; Wilson, B. R.; Wilson, A. K. *J. Chem. Phys.* **2007**, *127*, 124110.

(39) Mintz, B.; Friskell, S.; Shah, A.; Wilson, A. K. *Int. J. Quantum Chem.* **2007**, *107*, 3077.

(40) Rienstra-Kiracofe, J. C.; Tsuchumper, G. S.; Schaefer, H. F., III.; Nanadi, S.; Ellison, G. B. *Chem. Rev.* **2002**, *102*, 1231.

(41) Jensen, F. *J. Chem. Phys.* **2002**, *117*, 9234.

CT800575Z

JCTC Journal of Chemical Theory and Computation

Mobile Block Hessian Approach with Adjoined Blocks: An Efficient Approach for the Calculation of Frequencies in Macromolecules

A. Ghysels,[†] V. Van Speybroeck,[†] E. Pauwels,[†] D. Van Neck,[†] B. R. Brooks,[‡] and M. Waroquier*,[†]

Center for Molecular Modeling, Ghent University, Proeftuinstraat 86, B-9000 Gent, Belgium, and Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892

Received November 11, 2008

Abstract: In an earlier work, the authors developed a new method, the mobile block Hessian (MBH) approach, to accurately calculate vibrational modes for partially optimized molecular structures [*J. Chem. Phys.* 2007, *126* (22), 224102.]. It is based on the introduction of blocks, consisting of groups of atoms, that can move as rigid bodies. The internal geometry of the blocks need not correspond to an overall optimization state of the total molecular structure. The standard MBH approach considers free blocks with six degrees of freedom. In the extended MBH approach introduced herein, the blocks can be connected by one or two adjoining atoms, which further reduces the number of degrees of freedom. The new approach paves the way for the normal-mode analysis of biomolecules such as proteins. It rests on the hypothesis that low-frequency modes of proteins can be described as pure rigid-body motions of blocks of consecutive amino acid residues. The method is validated for a series of small molecules and further applied to alanine dipeptide as a prototype to describe vibrational interactions between two peptide units; to crambin, a small protein with 46 amino acid residues; and to ICE/caspase-1, which contains 518 amino acid residues.

I. Introduction

Conformational changes of macromolecules can be probed by a variety of experimental techniques such as X-ray crystallography, NMR spectroscopy, and so on.[1−4] The theoretical prediction of conformational flexibility is far from straightforward, especially for very large molecules such as proteins. In many cases, collective motions are present that occur on a time scale that is too long (on the order of milliseconds) to be accessible through molecular dynamics simulations.[5,6] Normal mode analysis (NMA) has proven successful in representing domain and hinge-bending motions in proteins, and it has been shown that, for several systems, the lowest-frequency modes contribute the most to conformational changes.

Until recently, theoretical studies of dynamical properties by NMA were restricted to rather small proteins.[7] This is partly due to the fact that a standard NMA needs to be preceded by an energy minimization. Because of the size of biological systems, this is already a computationally exhausting task. In addition, an NMA requires the diagonalization of a $3N_a \times 3N_a$ matrix (where N_a is the number of atoms). As the system becomes larger, this too becomes computationally very expensive.

A series of models has been proposed to simplify both the potential energy function[8] and the normal mode equations. For a detailed description, we refer to a recent review by Tama.[9] In particular, the method introduced by Durand et al.[10] and later extensively tested by Tama et al.[11] on relatively large biosystems (up to citrate synthase with 856 residues) deserves attention. It makes use of the concept that a protein chain can be seen as a sequence of rigid components, i.e., the peptide units. The combination of rotation

* Corresponding author e-mail: Michel.Waroquier@UGent.be.
[†] Ghent University.
[‡] National Institutes of Health.

10.1021/ct800489r CCC: $40.75
Published on Web 03/30/2009

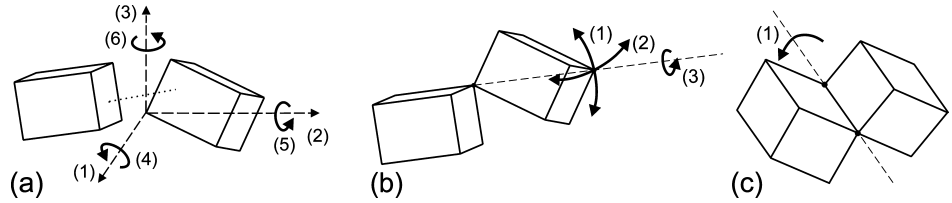


Figure 1. Degrees of freedom in the relative motion of (nonlinear) adjoined blocks. (a) Nonadjoined blocks have six degrees of freedom. The dotted line visualizes the bonding/nonbonding interaction between the blocks. (b) Blocks linked by one adjoining atom have three degrees of freedom. (c) Blocks linked by two adjoining atoms (hinge-type connection) have one degree of freedom.

and translation motions of these rigid blocks already gives a fair description of the lowest-frequency modes. For that reason, Tama et al. referred to this method as the RTB method, which stands for Rotations−Translations of Blocks.[11] In the approach developed by Durand et al.,[10] the possibility exists to obtain, in a second step, the exact normal modes by an iterative procedure, using effective Hamiltonians that couple the RTB modes with the higher-frequency internal modes of each block. This approach, however, might be problematic for larger systems, as one deals with a highly degenerate low-frequency spectrum in this case. Tama et al. showed that a good approximation of the lowest-frequency spectrum could already be obtained by performing only the first step.

Recently, we developed the mobile block Hessian (MBH) approach, which, in a sense, very much resembles the RTB method developed by Durand et al., as each block is allowed to move with six degrees of freedom attributed to translation and rotation.[12−14] The main difference between our original MBH method and the RTB method lies in the fact that the MBH method can also handle partially optimized systems. This kind of system also frequently occurs in other application fields, such as reactions taking place in porous materials or on surfaces. In that case, only part of the system is optimized, i.e., the chemically active part, whereas the rest is kept fixed to prevent unphysical deformations of the molecular environment. As a consequence of the partial optimization, the Hessian will have only three zero eigenvalues instead of six, implying that the rotational invariance of the potential surface is no longer manifest. Moreover, a set of spurious imaginary frequencies might appear.[12] In 2002, Li and Jensen introduced the partial Hessian vibrational analysis (PHVA) method, in which the fixed atoms are given an infinite mass during the frequency calculation and thus can no longer participate in the small-amplitude vibrations.[15] This method was successfully applied by Besley and Metcalf to calculate the amide I band of polypeptides and proteins.[16] More sophisticated schemes were introduced by Head and co-workers that allowed coupling between PHVA modes and modes in the fixed part.[17−20]

In this work, we introduce an extension of our original MBH method, in which the blocks can be linked to each other by one or two common atoms (adjoining atoms), as schematically shown in Figure 1. This is important, for instance, in very large (bio)systems, where one would like to have the possibility of calculating specific normal modes at decreased computational cost, as the number of degrees of freedom is further reduced by the introduction of adjoining atoms. If the blocks are linked by one atom, the block has three rotational degrees of freedom with respect to a previous block. If the blocks are linked by two atoms (creating a hinge-type connection), only one degree of freedom for the connected block remains. The linkage between blocks is a fundamental difference from the RTB method. Moreover, the MBH method is valid for partially optimized systems, even in the present adjoined version, extending the possible applications.

The newly introduced method is first tested on some basic test examples such as dimethylether and isobutane by assessing whether the system with reduced dimensionality can reproduce the original benchmark frequencies and associated normal modes. The hinge-type connection is tested for a series of *n*-alkanes and cyclohexane. Furthermore, the method is validated on a frequently investigated test example, alanine dipeptide, containing two peptide bonds. Finally, both connection schemes for the blocks are tested for crambin, a small protein with 46 amino acids. This system serves as a good test case because the MBH results can be compared to other methods that have also been tested on this polypeptide. To illustrate the ability of the MBH approach to treat quite large systems, the method is applied to ICE/caspase-1 with 518 amino acids.

II. Theoretical and Computational Approach

A. MBH Partitioning of the System. We distinguish among three different partitioning schemes, as illustrated in Figure 1. In the original implementation of the MBH method, atoms could only be part of one MBH block (a). In the current MBH approach, atoms can be part of two or more rigid blocks. Such a common atom, i.e., an adjoining atom, plays the role of a joint between the blocks. Two basic linkages can be realized with adjoining atoms: blocks connected by one adjoining atom (b) and blocks connected by two adjoining atoms (c). Unlinked blocks have six degrees of freedom in their relative motion, but blocks connected by an adjoining atom have only three (the three translational degrees of freedom are lost). The remaining three degrees of freedom correspond to rotations about an axis through the adjoining atom. If blocks are connected by two adjoining atoms to a previous block, the linkage is a hinge-type connection. Only one degree of freedom remains, corresponding to a rotation about the axis through the two adjoining atoms.

The hinge-type connection can correspond to some typical internal motions in molecules, as illustrated in Figure 2. The

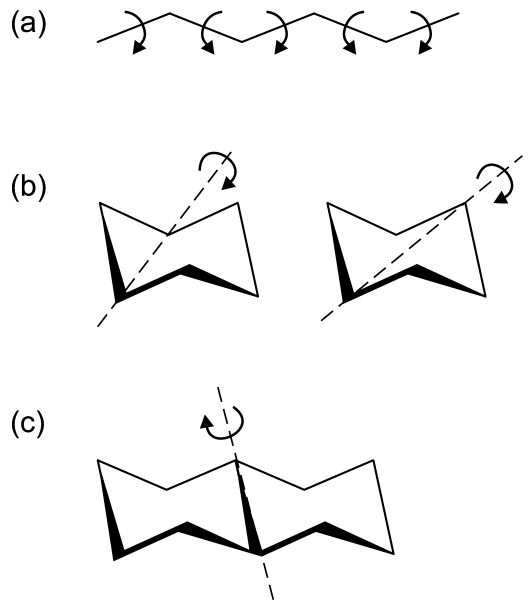


Figure 2. Three categories of motion characterized as involving hinge-type connections: (a) internal rotations, (b) ring inversion, (c) bending mode of two rings with respect to each other.

most important case (a) corresponds to internal rotations about a bond and occurs very frequently in chemistry. The hinge is determined by the atoms defining the rotational axis, i.e., by the chemical bond. Internal rotations in a molecule have frequently been studied in the literature, as they can show strong anharmonic motions.[21−27] The second case (b) can be found in ring compounds, in which ring inversions can take place. The hinge is defined in this case by two atoms that are not directly connected by a chemical bond. The third case (c) is found in fused ring systems where two rings can bend with respect to each other. Here, the hinge corresponds to the single bond that is common to the two rings.

B. Computational Details. Preference was given to the use of one single molecular mechanics program package, even for the small test molecules. The aim of the work was not the accurate reproduction of experimental frequencies for the relevant modes, but a validation of the MBH method with respect to full Hessian (benchmark) NMA. All calculations were performed with the PARAM22/27 (proteins) and PARAM34 (alkanes, ethers) force fields of the CHARMM package.[28] First, the geometry was fully optimized (root-mean square of the gradient lower than 10^{-6} kcal mol^{-1} Å^{-1}). For the dimethylether test case, two partially optimized geometries were generated as well using the SHAPE and restraints commands (RESD) from the CONS module. Next, the full Hessian was calculated in CHARMM and diagonalized with the VIBRAN module to obtain the benchmark (full Hessian) frequencies and modes.

C. Implementation. The extension of the MBH method to the case of adjoined blocks was implemented using a restraint technique and dummy atoms. This approach has the advantage that none of the MBH formulas in ref 13 need be changed. Only the Hessian must be complemented to impose the constraints. The procedure works for an arbitrary number of adjoining atoms connecting an arbitrary number of blocks,

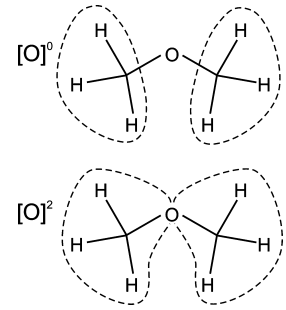


Figure 3. Dimethylether: Block choices.

so both linkages with one adjoining atom and linkages with two adjoining atoms can be treated.

Suppose an adjoining atom A belongs to s_A blocks. We can formally replace the adjoining atom by s_A dummy atom copies, A_1, ..., A_{s_A}, at the same reference position. The mass of the dummy atoms is the original mass m_A divided by s_A. Each of the dummy atoms is assigned to one of the blocks. In this way, a strict partition (no adjoining atoms anymore) of the atoms over the blocks is restored, and the former MBH computational routines can be used.

We must still impose the fact that the dummy copies, in reality, describe the same adjoining atom A, i.e., during the motion, the coordinates $\boldsymbol{r}_{A_1} = \cdots = \boldsymbol{r}_{A_{s_A}}$ should coincide. This can be done by adding to the potential energy harmonic terms of the form

$$V_A = \frac{1}{2}\kappa \sum_{i=2}^{s_A} |\boldsymbol{r}_{A_1} - \boldsymbol{r}_{A_i}|^2 \qquad (1)$$

for each adjoining atom. Choosing the spring constant κ sufficiently large will ensure that, during the motion, the positions of the dummy copies (nearly) coincide; otherwise, a large energy penalty is generated. At the level of the normal modes, the introduction of the dummy copies will lead to more vibrations than in the true system. However, the addition of the V_A terms to the potential energy will result in a clear energy separation between the physical modes where the dummy atoms coincide and the unphysical modes involving stretches between dummy copies of the same atom. These unphysical modes are all at much higher energies and can be easily deleted from the frequency list.

The value of κ to be used depends on the following considerations: On one hand, κ must be in the decoupling regime, so that the frequency spectrum of the physical modes does not depend on its value. On the other hand, an exageratedly large value for κ can lead to numerical instability as the condition number of the Hessian matrix deteriorates. In practice, a value of about 10^7 hartree/(Bohr length)^2 was found to provide excellent results.

In the following discussion, the share number s_A is used to label the different block choices. For instance, in Figure 3, $[O]^2$ specifies the block choice where the oxygen atom belongs to two blocks.

III. Results and Discussion

A. Small Test Molecules for Blocks Linked by One Adjoining Atom. *1. Dimethylether.* A first test case is

Table 1. Dimethylether: Frequency Calculations with the Fully Optimized Structure and Two Partially Optimized Structures[a,b]

fully optimized					partially optimized $[O]^0$		partially optimized $[O]^2$	
	blocks					blocks		blocks
full	$[O]^0$		$[O]^2$		full	$[O]^0$	full	$[O]^2$
0	0		0		−28	0	0	0
0	0		0		0	0	0	0
0	0		0		0	0	0	0
0	0		0		0	0	59	0
0	0		0		12	0	156	0
0	0		0		58	0	193	0
177	177	(100)	178	(100)	199	176	250	175
244	245	(100)	245	(100)	257	248	323	232
478	479	(100)	497	(98)	482	483	496	480
898	931	(97)			897	932	975	
1023	1061	(94)			1027	1061	1111	
1096	1135	(91)			1094	1136	1118	
1149	1201	(86)			1145	1203	1363	
1149	1209	(86)			1156	1212	1405	
1204	1244	(87)			1216	1247	1430	
1437					1443		1449	
1442					1445		1471	
1442					1450		1576	
1460					1463		2099	
1583					1571		2100	
1597					1585		2176	
2850					2848		2849	
2850					2849		2914	
2912					2912		2916	
2913					2913		5423	
2915					2915		5434	
2916					2919		5644	

[a] For each structure, the full Hessian (full) frequencies are compared with the MBH (blocks) frequencies, in cm^{-1}, including the frequencies due to global translation/rotation. [b] Maximum square overlap, $|\langle \nu^{MBH}|\nu^{bench}\rangle|^2$, given in parentheses (in %).

dimethylether. The full Hessian spectrum of the fully optimized structure has 21 normal modes in addition to the six global translational and rotational modes and serves as the benchmark (Table 1). The three lowest modes of dimethylether correspond to rotations of the methyl groups about an axis through the oxygen atom. Two block choices are considered as depicted in Figure 3. A natural choice is to put the methyl groups in two blocks, whereas the O atom in the center is a free atom not belonging to any block: $[O]^0$. This block choice corresponds to the original MBH approach. The number of degrees of freedom can be further reduced by omitting the stretch between the oxygen atom and the methyl groups. This is realized in block choice $[O]^2$, where the oxygen atom is taken up in both blocks. Table 1 lists the MBH frequencies, with the maximum square overlap with the benchmark, $|\langle \nu^{MBH}|\nu^{bench}\rangle|^2$, given in parentheses.

The results from the full optimization are discussed first. The MBH frequencies for block choice $[O]^0$ lie close to the benchmark, and the overlap values are larger than 86%. Only nine intrinsic normal modes remain, originating from the six degrees of freedom for each block, plus three translational degrees of freedom for the oxygen, minus the six global rotational and translational degrees of freedom. In the case of $[O]^2$, the oxygen atom acts as an adjoining atom between the two blocks. As a consequence, the C—O stretches no longer appear in the spectrum. Only three modes remain, which describe exactly the lowest normal modes with excellent overlaps of 98−100%. The value of the third lowest frequency has increased (497 cm^{-1}) because adjoining blocks increase the stiffness of the system, thereby giving rise to higher frequencies.

The last four columns of Table 1 show frequencies calculated for partially optimized structures. These were obtained in the following way: First, an initial nonequilibrium geometry was generated by disturbing the positions of some of the atoms by 0.1 Å from their global equilibrium positions. After the blocks had been defined (in this specific application, we considered $[O]^0$ and $[O]^2$), an energy minimization was performed with the internal coordinates within each block frozen. This led to a partially optimized structure. The individual atoms of the blocks might feel residual forces, but the positions and orientations of the blocks are optimized with respect to each other. For the thus-obtained partially optimized structure (which is different for each specific block choice), both a full Hessian calculation (column labeled "full") and an MBH calculation (column labeled "blocks") were carried out.

Partial optimization reduces the computation time for the geometry optimization; however, a standard full Hessian calculation on partially optimized structures might yield spurious low or even imaginary frequencies (indicated as negative values in Table 1). Also, unrealistically high frequencies can be detected in the vibrational spectrum (ca. 5000 cm^{-1}). Application of the MBH avoids such unphysical results. Moreover, the MBH frequencies from the partially optimized structures closely resemble those from the fully optimized structure, as can be seen in Table 1. The present MBH approach is therefore well-suited to treat partially

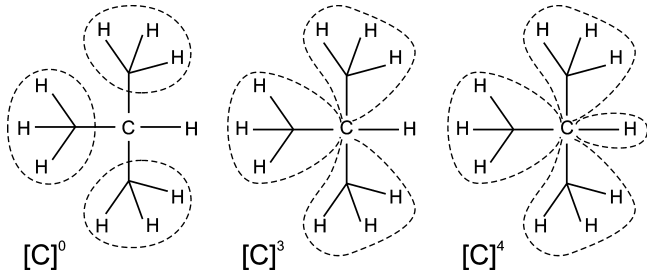


Figure 4. Isobutane: Block choices.

Table 2. Isobutane: Comparison of Benchmark (Full Hessian) and MBH Frequencies (cm^{-1}) for Different Block Choices[a]

	blocks					
full	$[C]^0$		$[C]^3$		$[C]^4$	
248	248	(100)	248	(100)	248	(100)
263×2	263	(100)	263	(100)	263	(100)
361×2	363	(100)	373	(99)	373	(99)
428	430	(100)	456	(97)	458	(97)
795	800	(99)				
917×2	928	(98)				
959×2	977	(97)				
961	978	(97)				
1117	1133	(96)				
1139×2	1149	(97)				
1356×2	1364	(89)	1258	(51)	1258	(51)
1399×2						
1400						
1426						
1429×2						
1435×2						
1436						
2901×3						
2944	2947	(79)	2916	(77)		
2957						
2958×2						
2960×2						
2963						

[a] Maximum (summed if degenerate modes) square overlap, $|\langle \nu^{MBH}|\nu^{bench}\rangle|^2$, given in parentheses (in %).

optimized structures. This simple test molecule illustrates that well-chosen adjoined blocks provide an economical way to reproduce interesting low-frequency normal modes.

2. Isobutane. As a second test case, we consider isobutane, where one can choose multiple adjoined blocks, all connected through the central carbon atom that serves as the adjoining atom. All investigated block choices are depicted in Figure 4. Case $[C]^0$ corresponds to nonadjoined blocks, whereas cases $[C]^3$ and $[C]^4$ include linkages. Table 2 compares the MBH results with the benchmark (full Hessian) spectrum. Isobutane has 12 degenerate benchmark frequencies because of the symmetry of the molecule. When calculating the maximum square overlap, we have taken into account this degeneracy by summing the square overlaps, $\sum_i |\langle \nu^{MBH}|\nu_i^{bench}\rangle|^2$, over the modes $|\nu_i^{bench}\rangle$ that belong to the same degenerate frequency ν^{bench} and picking the maximum summed square overlap.

Case $[C]^0$ with three methyl groups figuring as nonadjoined blocks yields good correspondence of the frequencies and also very good overlap of the modes (79% and higher). This is in full accordance with the previous MBH results reported in ref 12.

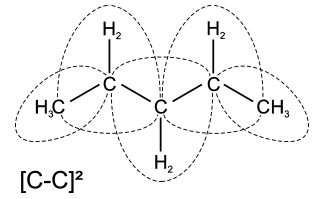


Figure 5. *n*-Alkanes: Block choice for pentane.

In the case of $[C]^3$, where the three methyl blocks are linked to each other by including the C atom in each of the three methyl blocks, the four lowest-frequency modes are very well reproduced by 97−100%. Some frequencies have increased with respect to the benchmark values because of the stiffening of the system. The remaining frequencies in the medium- and high-frequency spectrum, i.e., 1258 and 2916 cm^{-1}, correspond to motions of the free H atom (not belonging to any of the blocks): bending with respect to the three methyl groups and stretching of the C—H bond. Especially for the bending motion, the overlap is reduced, as this motion is expected to couple with the internal motions in the methyl block (steric hindrance hydrogens).

Case $[C]^4$ groups the H atom that was previously a free atom with the C atom into an additional linear adjoined block. Consequently, the C—H stretch is eliminated, whereas the rest of the frequencies and overlaps remain virtually unchanged. This illustrates the fact that the high-frequency stretch is not coupled to the lower-frequency spectrum. Such adiabatic behavior is the main justification for the use of constraints for the X—H bonds (N—H, O—H, etc.) in molecular dynamics simulations of extended (bio)systems (e.g., the SHAKE algorithm implemented in CHARMM[29]). The C—H stretches occur at a much shorter time scale and are not essential for configurational changes and other characteristics. Similarly, the C—H stretches are not relevant for the lower spectrum of the NMA, thereby justifying block choice $[C]^4$.

B. Small Test Molecules for Blocks Linked by Two Adjoining Atoms. *1. n-Alkanes.* Internal rotations belong to the category of hinge-type connections. We illustrate this by considering subsequent internal rotations in *n*-alkanes, going from propane to decane. Figure 5 shows how each $C—CH_2—C$ unit forms a block, such that subsequent blocks always have the C—C bond in common. The blocks at the ends of the chain consist of a $C—CH_3$ unit. This block choice can be denoted by $[C—C]^2$, pointing out that each C—C bond belongs to two blocks. The only degrees of freedom that are not fixed by the partitioning are the dihedral angles of the internal rotations about the C—C bonds (see also Figure 2a).

Table 3 compares the MBH frequencies to the lowest full Hessian frequencies, which serve as the benchmark. The number of frequencies is reduced substantially from $9n$ (full) to only $n - 1$ frequencies, for a C_nH_{2n+2} alkane. Still, the MBH frequencies resemble the benchmark fairly well, and only a slight frequency enhancement is noticed, despite the rather drastic approximation. The overlap between MBH and full modes is 99% and higher, which means the lower spectrum of the *n*-alkanes is indeed built up of dihedral angle bending motions. Several gaps are noticed in the spectrum,

Table 3. Alkanes: Benchmark (Full Hessian) Frequencies[a] Compared to MBH Frequencies (cm^{-1})[b]

propane			butane			pentane			hexane		
full	[C—C]^2		full	[C—C]^2		full	[C—C]^2		full	[C—C]^2	
225	226	(100)	126	131	(100)	115	125	(100)	78	86	(99)
258	265	(99)	222	231	(99)	128	129	(100)	110	114	(99)
376			258	261	(100)	206			155		
⋯			293			244	249	(99)	161	167	(99)
			395			246	250	(100)	239	244	(99)
			⋯			356			249	254	(100)
						⋯			290		
									⋯		

heptane			octane			nonane			decane		
full	[C—C]^2		full	[C—C]^2		full	[C—C]^2		full	[C—C]^2	
61	69	(100)	46	52	(100)	38	43	(100)	30	35	(100)
89	91	(100)	79	81	(100)	68	70	(100)	62		
118			93			75			63	64	(100)
142	151	(99)	113	123	(99)	96	105	(99)	78	87	(100)
161	164	(100)	149	153	(99)	131	133	(100)	121	123	(100)
244	249	(99)	168	174	(100)	155	164	(99)	133	142	(99)
245	249	(100)	217			171	175	(100)	155		
265			243	248	(99)	187			165	169	(99)
277			246	250	(100)	210			171	177	(100)
⋯			246			244	244	(99)	193		
			363			245	250	(100)	244	244	(99)
			⋯			322			245	249	(100)
						⋯			272		
									⋯		

[a] Only lowest listed. [b] Maximum square overlap, $|\langle \nu^{MBH} | \nu^{bench} \rangle|^2$, given in parentheses (in %).

e.g., the third frequency of pentane (206 cm^{-1}) is missing, but visualization of this mode reveals that it mainly involves C—C—C angle bending, which is impossible to describe with the [C—C]^2 block choice. Summarizing, the study of *n*-alkanes demonstrates that the MBH approach with hinges captures the low-frequency modes associated with dihedral angle bending.

2. Cyclohexane. The hinge-type connection can also be used to select the rotation about an axis that does not coincide with a bond, as shown in Figure 2b, for instance, to describe puckering modes. This application is tested on the chair conformation of cyclohexane, which contains in the lower vibrational spectrum six ring bending modes at 232, 375, 439, and 547 cm^{-1}. The 232 and 438 cm^{-1} modes are doubly degenerate. These vibrational modes are schematically depicted in Figure 6a, for the hypothetical case that the molecule were planar. Modes 1—3 correspond to out-of-plane modes, whereas modes 4—6 correspond to (mainly) in-plane modes.[30] For the realistic nonplanar chair conformation of cyclohexane, the same terminology will be used but referring to the plane perpendicular to the 3-fold axis of the molecule. Various block choices were tested to investigate the extent to which the ring vibrations can be reproduced by the MBH approach (see Figure 6b). In the first case, [C1—C4]^2, the carbon ring is spit into two adjoined blocks using the hinge-type connection. The blocks can rotate with respect to each other about the common C1—C4 axis. The C—H stretches are eliminated by considering each C—H pair as a linear block connected to the ring. Equivalent splittings (because of the symmetry) have also been considered, i.e., [C2—C5]^2 and [C3—C6]^2. In the second case, [C1—C3]^2 + [C3—C5]^2 + [C5—C1]^2, the ring is split in three blocks, subsequently

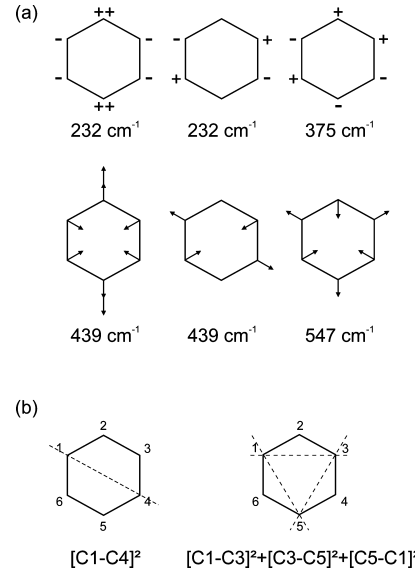


Figure 6. Planar representation of cyclohexane. (a) Six lowest-frequency bending modes. The first three are out-of-plane motions (±), and the next three are in-plane motions (arrows). (b) Block choices.

connected by one adjoining atom. This special case forms a ring structure of adjoined blocks.

The doubly degenerate mode at 439 cm^{-1} can never be reproduced accurately with our block choices (maximum square overlap of no more than 16%) because these correspond to C—C—C angle bendings. From the schematic representation of the modes in Figure 6, it can be anticipated that the last block choice is most suited to describe the 375 and 547 cm^{-1} modes. Both modes correspond to the flattening of the carbon ring. In the 375 cm^{-1} mode,

alternating CH_2 units rotate in opposite directions perpendicular to the plane. For the 547 cm^{-1} mode, the CH_2 units move in phase, and some C—C—C angle bending is involved. These modes can be best reproduced if the ring is split into three adjoined blocks (84% and 60%, respectively). One of the doubly degenerate 232 cm^{-1} modes coincides with the folding double motion and is indeed reproduced with the first block choice $[C1—C4]^2$ by 97%. The other 232 cm^{-1} mode represents a twist of the ring, which is a superposition of two folding double motions. Hence, the first block choice can only partially reproduce this mode (maximum square overlap of 74%). Splitting the ring into three blocks gives a better estimate of the mode, with a maximum square overlap of 84%.

This analysis illustrates how the nature of the complicated ring normal modes can be investigated with the MBH approach by examining the square overlaps. This is a useful feature, especially when the nature of the modes cannot be deduced uniquely by visualization. For the cyclohexane test case, it allowed us to confirm the nature of the six lowest bending modes.

C. Peptide Chains. *1. Block Choices.* The next step is to investigate whether the adjoined-blocks MBH model is appropriate for simulating the low-frequency modes of peptide chains. The peptide bond is known to be a stable planar unit, and it is straightforward to introduce blocks that include the entire peptide bond. Peptides have the ability to form different secondary structure elements as a result of the conformational flexibility of the backbone. There are essentially two backbone degrees of freedom for the amino acid residue: the dihedral angles ϕ (C—N—C_α—C) and ψ (N—C_α—C—N). The angles describe rotations about the N—C_α and C_α—C bonds, respectively.

Figure 7 provides an overview of the considered plausible partitioning schemes for peptide chains. The smallest block choice corresponds to considering consecutive peptide bonds (C═O—NH) as rigid blocks, labeled by $[C_\alpha]^0$. In case $[C_\alpha]^1$, a complete residue—peptide bond plus side chain—is considered as a block. This corresponds to the RTB approach introduced by Durand en al. and extensively tested by Tama et al. on a large variety of proteins.[10,11] From a methodological point of view, it corresponds to the MBH approach introduced in our earlier articles.[12−14] It is also possible to combine the adjoined blocks with atoms not included in any block. This is useful for the calculation of localized modes of interest, e.g., in the chemically active part. Such block choices have not been considered in this work.

The adjoined MBH approach allows the various blocks to be connected by one or more adjoining atoms. As a first approximation, the peptide units can be linked by the C_α atoms. This further reduces, by three, the total number of degrees of freedom for each considered block. Several options are possible for treating the side chain and the H atom connected to the C_α carbon, such as $[C_\alpha]^2$, $[C_\alpha]^{3a}$, $[C_\alpha]^4$, and $[C_\alpha]^{3b}$ in Figure 7, which will be discussed later with respect to the alanine dipeptide test case. Note that the actual block choice for the ends of the chain has little influence for long peptide chains.

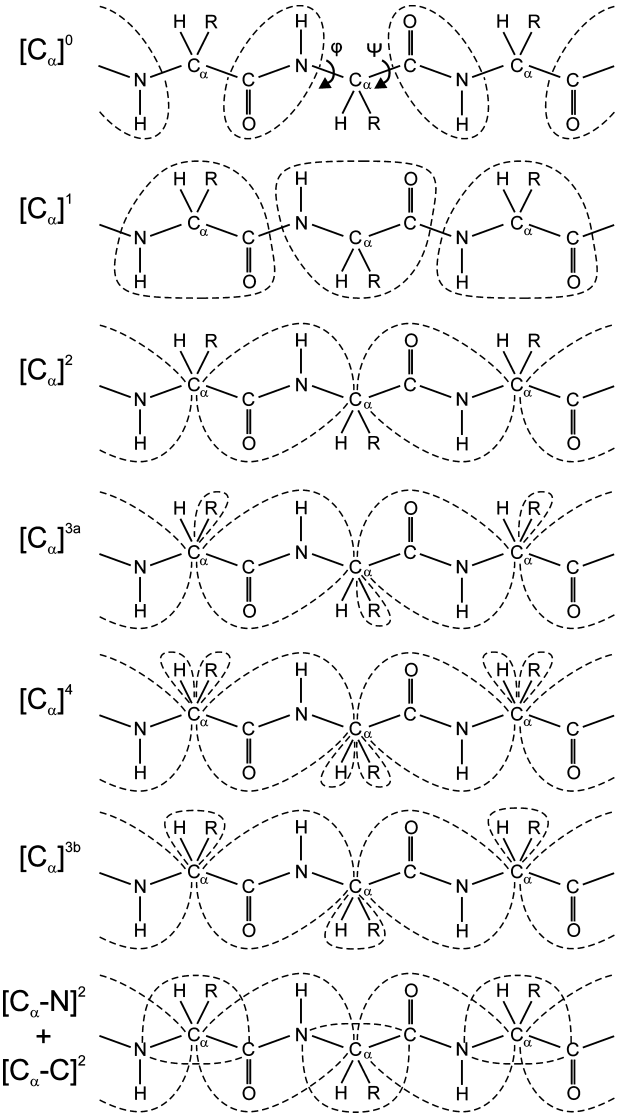


Figure 7. Partitioning schemes for peptide chains. Nonadjoined blocks: $[C_\alpha]^0$ and $[C_\alpha]^1$. Blocks linked by one adjoining atom: $[C_\alpha]^2$, $[C_\alpha]^{3a}$, $[C_\alpha]^4$, $[C_\alpha]^{3b}$. Blocks linked by two adjoining atoms: $[C_\alpha—N]^2 + [C_\alpha—C]^2$. Dihedral angles ϕ and ψ representing the two backbone degrees of freedom are also shown for the top chain.

Finally, the subsequent blocks can share two adjoining atoms as shown at the bottom of Figure 7. Each introduced block shares the C_α—N bond or the C_α—C with the block connecting the subsequent peptide units. As a result, the only degrees of freedom that are left over are the internal rotations about the common C_α—N and C_α—C bonds, which are exactly the ϕ and ψ dihedral angles. This block choice is labeled by $[C_\alpha—N]^2 + [C_\alpha—C]^2$, referring to the common C_α—N and C_α—C bonds.

By simply choosing adjoined blocks, as in $[C_\alpha—N]^2 + [C_\alpha—C]^2$, one is able to select the dihedral angles as variables. The MBH model is thus useful for NMA with variables of interest, without the need for extra computational implementations for the construction of the variables (i.e., no complicated transformation to internal coordinates is required). As for the geometry optimization, only the ϕ and ψ dihedral angles have to be force-free; other internal coordinates do not necessarily have to be optimized in the MBH

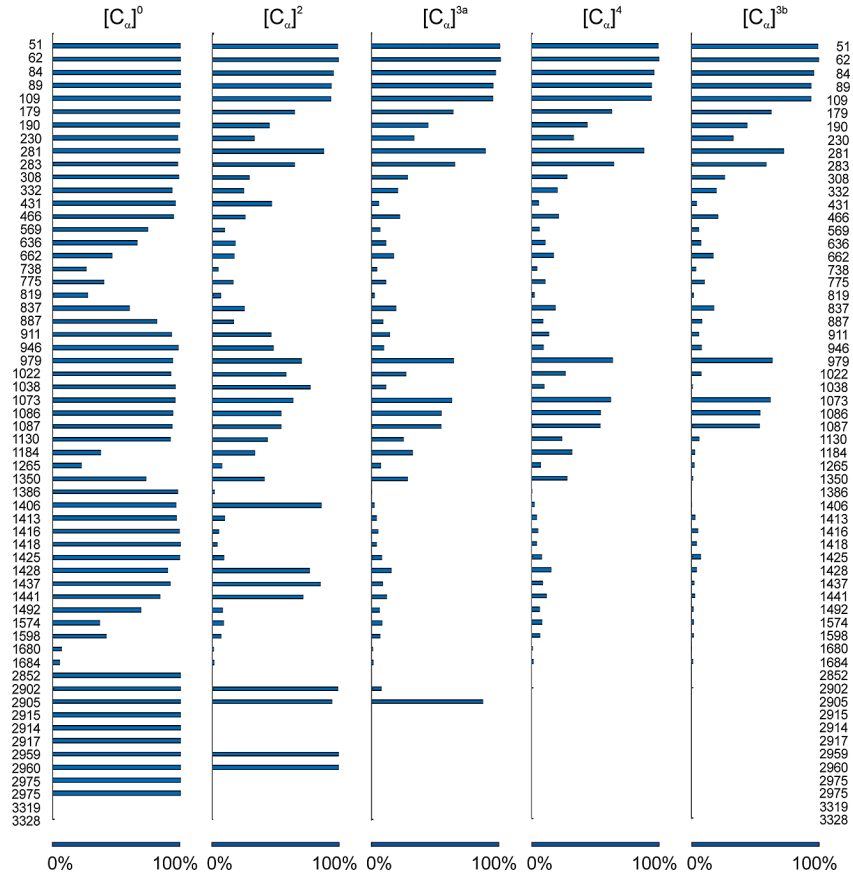


Figure 8. Alanine dipeptide: Cumulative square overlap P_j (%) for each non-zero-frequency benchmark mode.

framework. From a computational point of view, the reduction of the Hessian size is impressive. Only two degrees of freedom (ϕ, ψ) are considered per residue. A protein with N_R residues easily has over $30N_R$ degrees of freedom, but the adjoined blocks with hinges reduces this number to $2N_R$.

2. Alanine Dipeptide. Blocked (or capped) alanine dipeptide (N-acetyl-L-alanine-N′-methylamide) is commonly studied as a prototype of nonglycine/nonproline protein backbones, since it allows full sampling of the ϕ/ψ conformational space without the additonal complexity of side-chain degrees of freedom.[31] Numerous computational and experimental studies of alanine dipeptide have explored the thermodynamic,[32,33] kinetic[34−36] and spectroscopic properties.[37,38] Alanine dipeptide takes several conformations in gas phase.[39,40] In this study, the stable conformer with $\phi = -81°$ and $\psi = 70°$, usually referred to as $C7_{eq}$, was used.

To quantify how a benchmark normal mode, resulting from a full NMA analysis, corresponds to an MBH normal mode, the square overlap between the corresponding eigenvectors is again used. However, as the system size increases, many modes become degenerate or nearly degenerate. It is then more appropriate to use a cumulative square overlap P_j, defined as

$$P_j = \sum_i |\langle \nu_i^{\mathrm{MBH}} | \nu_j^{\mathrm{bench}} \rangle|^2 \qquad (2)$$

which quantifies how well the benchmark mode $|\nu_j^{\mathrm{bench}}\rangle$ can be reproduced by all MBH modes $\{|\nu_i^{\mathrm{MBH}}\rangle\}$. The overlap does not necessarily become 100% because the MBH modes do not form a complete basis.

The efficiency and adequacy of the adjoined MBH method are best illustrated by examining the cumulative square overlap P_j for each non-zero-frequency benchmark mode, as can be seen in Figure 8. The block choices $[C_\alpha]^2$, $[C_\alpha]^{3a}$, $[C_\alpha]^4$, and $[C_\alpha]^{3b}$ from Figure 7 are considered to investigate how the side chain should be treated. To allow a full comparative study the results of block choice $[C_\alpha]^0$, where the peptide units are considered as nonadjoined blocks, are also shown in the first column of Figure 8. The 14 lowest-frequency modes and the high-frequency C—H stretches are reproduced by the MBH modes for at least 97%. Only some of the modes in the medium-frequency range are less well reproduced.

In the next four cases, blocks are linked to each other with the C_α atom as the adjoining atom. The ending methyl groups are also considered as blocks adjoined to the peptide unit blocks. In the case of $[C_\alpha]^2$, the side-chain atoms and the hydrogen atom are still free atoms. The frequencies are overestimated as a result of the stiffening of the system, but the cumulative square overlap indicates that the five lowest benchmark modes are still more than 90% reproduced. Case $[C_\alpha]^{3a}$ further reduces the number of frequencies by considering the side chain as an adjoined block as well (only the hydrogen atom is a free atom). The overlap in the lower spectrum is almost not influenced.

In case $[C_\alpha]^4$, the C_α—H stretch is eliminated by grouping the C_α and hydrogen atom in yet another adjoined block, such that the C_α atom links four blocks. As expected from the isobutane discussion, this C_α—H stretch is not coupled to the rest of the spectrum, which remains unchanged. Fixing

the angle between the C_α—H block and the side-chain block leads to case $[C_\alpha]^{3b}$ where the H atom is added to the side-chain block. Results largely resemble the previous $[C_\alpha]^4$ case, except for some medium frequencies around ± 1100 cm^{-1} associated with the side chain-C_α—H bending.

This example shows that grouping the atoms of the side chain into an adjoined block is a suitable choice to reproduce the lowest-frequency modes. Moreover the side-chain C_α—H bending and the C_α—H stretch are irrelevant for the lower spectrum, and the hydrogen atom can be included in the side-chain block (block choice $[C_\alpha]^{3b}$). The MBH method with linked blocks therefore appears to be an appropriate analysis tool for the lower spectrum of peptide chains.

3. Crambin. The MBH method with adjoined blocks is next applied to a longer, more representative peptide chain with a variety of amino acids. As an example, the well-known protein crambin (Protein Data Bank 1CCN) is chosen. The crambin molecule contains three disulfide bonds and displays β-strand, β-turn, and helical elements of protein secondary structure. The structure of crambin in water as determined by 2D NMR spectroscopy[41] was taken as a starting point for the geometry optimization (in the absence of the solvent). Crambin has 46 residues, including 648 atoms, and hence, the full NMA would result in 1944 frequencies. It is our aim to reduce the number of frequencies drastically by performing an MBH analysis on the system, but still to preserve the essential lowest-frequency normal modes.

The first block choice $[C_\alpha]^0$ considers the peptide groups (C═O—NH) as nonadjoined blocks, which is realizable with the previous MBH implementation. The square overlaps (not shown) revealed that the peptide units indeed move as nearly rigid groups, even when some of the residues are heavier and the peptide chain is longer.

Subsequently, it is investigated whether blocks can be adjoined. Three block choices from the list in Figure 7 are proposed, of which the first already gave promising results in the previous section: (1) case $[C_\alpha]^{3b}$ groups the side chain and the H atom in an adjoined block, (2) case $[C_\alpha]^1$ is the scheme introduced by Tama et al., and (3) case $[C_\alpha—N]^2 + [C_\alpha—C]^2$ uses the hinge-type connection.

Figure 9 compares the low-frequency part of the benchmark (full Hessian) spectrum with the MBH results. A grayscale indicates the square overlap between the modes. A dot on the diagonal indicates an exact frequency estimation, and the darker the dot, the better the overlap. A huge number of frequencies (1944) lies in a limited frequency interval—between 0 and 4000 cm^{-1}—and therefore, the square overlaps seldomly reach high values because of degeneracy. The cumulative square overlap P_j does not depend on the degeneracy as it sums over all MBH frequencies. Figure 10 shows a plot of the values of P_j for the benchmark modes below 50 cm^{-1}. It indicates how well a benchmark mode is still represented by all MBH modes.

Block choice $[C_\alpha]^{3b}$ overestimates the frequencies because of the stiffening of the system, but the cumulative overlap shows that reproduction of the benchmark modes is very good in the low-frequency range: all but one of the benchmark modes below 50 cm^{-1} are reproduced for at least 90% by the MBH modes, which is largely sufficient. Block

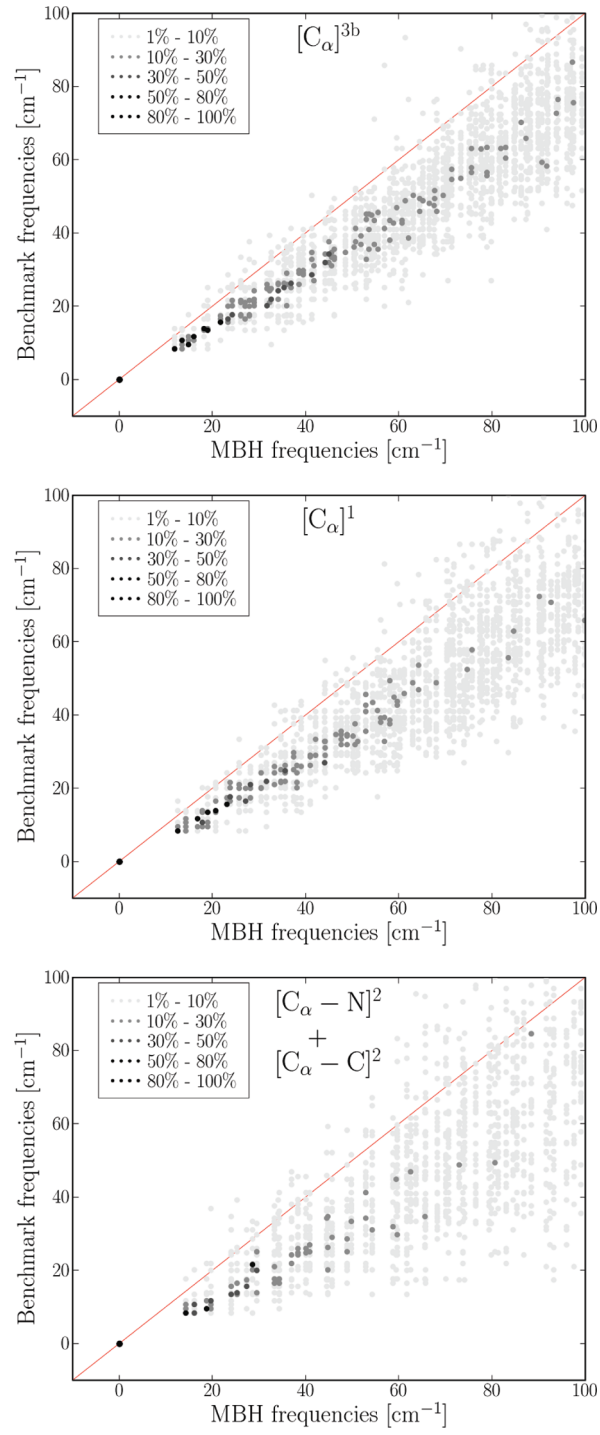


Figure 9. Crambin: Square overlap (%) between MBH modes and benchmark (full Hessian) modes in the 0−100 cm^{-1} frequency range for block choices $[C_\alpha]^{3b}$, $[C_\alpha]^1$, and $[C_\alpha—N]^2 + [C_\alpha—C]^2$.

choice $[C_\alpha]^{3b}$ differs from block choice $[C_\alpha]^4$ in the treatment of the H atom connected to the C_α carbon. Eliminating the C_α—H stretch without constraining the side chain C_α—H angle gives a somewhat more flexible model of the backbone, and case $[C_\alpha]^4$ is expected to be a more accurate approximation. However, the effect is minimal in the lower spectrum and becomes manifest only in the range above 250 cm^{-1} (not shown), because the H-atom motions interact very little with the backbone modes. Hence, when one is interested in the lowest frequencies, the side chain C_α—H angle bending

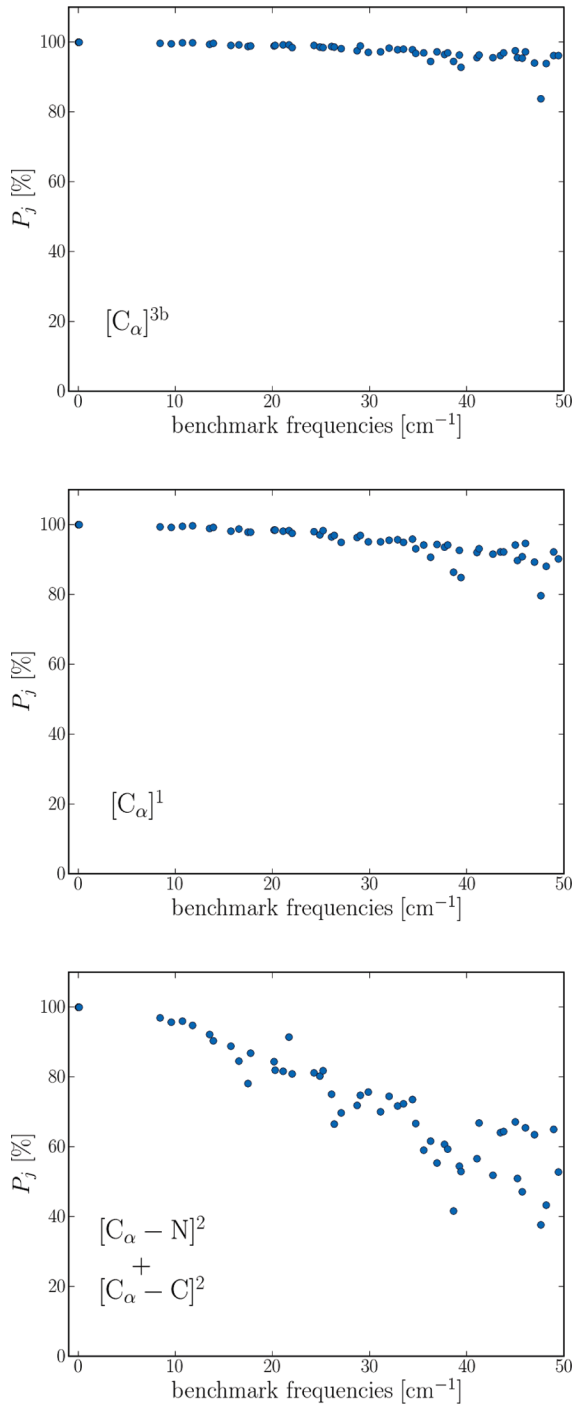


Figure 10. Crambin: Cumulative square overlap P_j (%) for benchmark modes in the 0−50 cm^{-1} range for block choices $[C_\alpha]^{3b}$, $[C_\alpha]^1$, and $[C_\alpha{-}N]^2 + [C_\alpha{-}C]^2$.

motion is not essential, and a partitioning as in case $[C_\alpha]^{3b}$, where all motions of the hydrogen atom are omitted, is sufficient.

The introduction of rigid blocks for the peptide chain implies a serious reduction of the number of degrees of freedom. In general, a peptide chain with N_{res} residues can be partitioned according to the $[C_\alpha]^{3b}$ block choice in a chain of N_{res} peptide blocks connected by the C_α atoms and N_{res} side-chain blocks linked to the chain by the C_α atoms as well. With three degrees of freedom per block, the total number becomes $6N_{res}$. For example, for crambin, the MBH method results in 276 modes, which is 14% of the initial 1944 variables.

A similar reduction was obtained by Tama et al.[11] with blocks consisting of several (n) peptide-plus-side-chain units. The blocks are not linked to each other, leading to a total number of $6N_{res}/n$ degrees of freedom. For comparison, we reproduced the overlap results in the case of one single peptide-plus-side-chain unit per block ($n = 1$). This partitioning, block choice $[C_\alpha]^1$ in Figure 7, gives the same reduction of modes as the $[C_\alpha]^{3b}$ partitioning with adjoined blocks. The overlap and cumulative overlap of $[C_\alpha]^1$ also perform well, although somewhat worse than the MBH approach with adjoined blocks. Thus, for the same degree of simplification, the introduction of adjoined blocks leads to a slight improvement of the partitioning proposed by Tama et al.

Tama et al. observed a linear relationship between the benchmark and the MBH frequencies

$$\nu_{MBH} = d\nu_{bench} \tag{3}$$

The factor d was obtained by a linear least-squares fit when comparing the lowest, say, n_{low}, benchmark modes with the n_{low} lowest MBH modes. For cases $[C_\alpha]^{3b}$ and $[C_\alpha]^1$, we find $d = 1.41$ and 1.49, respectively, when the lowest modes (below 40 cm^{-1}) are taken into account. The overestimation of the frequencies, i.e., $d > 1$, is due to the stiffening of the system. Applying the block constraints limits the motions to a hypersurface on the potential energy surface, which has higher curvatures than the total potential energy surface. In other words, the constraints make the system stiffer because the motion of blocks as a whole is more hindered than it would be if the atoms in the block were allowed to relax. On the other hand, the inertia of the blocks can lower the frequencies, because a block moving as a whole has more inertia than the atoms moving individually. The frequency change is thus governed by the balance between stiffening and inertia effects. The systematic frequency overestimation indicates that, in the lower spectrum of peptide chains, the stiffening is the dominating effect for the considered block choices.

The number of modes can be further reduced by using the partitioning scheme with hinges. The dihedral basis $[C_\alpha{-}N]^2 + [C_\alpha{-}C]^2$ as illustrated in Figure 7 was applied to crambin and reduced the number of modes to 96, which is as low as 4.9% of the initial number of variables. The results are again compared with the benchmark (full Hessian) frequencies in Figures 9 and 10.

The overlap plot indicates that fewer modes are reproduced and the frequency overestimation increases: $d = 1.85$. Still, the cumulative overlap P_j shows that 8 modes are reproduced by over 90%, whereas 20 modes are reproduced by over 80%. Visualization of the modes shows that the lowest MBH modes indeed represent large-amplitude motions where large parts of the chain move as a whole. The cumulative overlap decays rapidly when the frequency increases. In practice, of course, one usually investigates only the motions in the lowest five frequency modes for the purpose of analyzing conformational changes. Hence, it is clear that the simplification of block choice $[C_\alpha{-}N]^2 + [C_\alpha{-}C]^2$ still captures the

Table 4. Caspase-1: Lowest 14 Nonzero Benchmark (Full Hessian) Frequencies Compared to the Lowest MBH Frequencies (cm^{-1}), by Means of the Square Overlap $|\langle \nu^{MBH}|\nu^{bench}\rangle|^2$ (%)[a,b]

	benchmark													
MBH	3.4	3.6	3.9	4.8	5.4	5.4	5.7	5.7	5.8	5.8	6.0	6.5	6.6	7.3
6.5	11		**62**	21										
7.3	**25**			**61**										
7.6		**84**												
8.8					**33**	**46**								
9.2							**38**				**30**	18		
9.4					12	11		**42**		11				
9.6									**63**					
10.2							**38**				**34**			
10.6													**42**	
11.0					24			13					11	
11.9									11			**36**		
12.1														**63**
12.4										**46**				
12.6								13					11	
cum. sq. overlap	93	100	97	98	99	100	99	99	99	99	100	99	99	99

[a] Square overlaps below 10% are not shown; square overlaps above 25% are highlighted to emphasize the diagonal. [b] Cumulative square overlap P_j (%) given for each benchmark frequency at the bottom of the table.

essentials of the lowest modes and can be used as an analysis tool that focuses on the lower spectrum.

4. Caspase-1. To demonstrate the ability to treat quite large systems, we will briefly discuss the case of the interleukin-1 β converting enzyme (ICE or caspase-1, Protein Data Bank 1ICE).[42] This protein is a member of the caspase family of enzymes, which play a crucial role in apoptosis or programmed cell death. More specifically, caspase-1 proteolytically processes the interleukin-1β precursor to its active state and, hence, contributes essentially to the immune response.[43]

Caspase-1 is a tetramer, consisting of 518 residues and totaling 8252 atoms. A full Hessian diagonalization for the optimized structure would yield 24756 frequencies at a prohibitive expense. MBH with adjoined blocks considerably reduces the complexity to a fraction of the computational cost (less than 15%). For the calculation of the MBH frequencies, we made use of a recently developed method based on analytical constraints.[44] This method leads to the same numerical results as the dummy atom method described in section II but is much more efficient for extended molecular systems. The adjoined block choice $[C_\alpha]^{3b}$ was selected in view of its good performance in the crambin case. To show the potential of our MBH method, comparison was made with the 20 lowest exact full Hessian frequencies and modes, calculated with the DIAG option of CHARMM.

The six lowest frequencies correspond to the global translational and rotational modes. The next 14 nonzero modes lie in a narrow frequency range of 3.4−7.3 cm^{-1}, and their square overlaps with the lowest MBH frequencies, $|\langle \nu^{MBH}|\nu^{bench}\rangle|^2$, are given in Table 4. Square overlaps below 10% are not shown; square overlaps above 25% are highlighted. Because of the high degeneracy of the modes, the modes seem to be mixed up or to have changed order, but the highest square overlaps nevertheless lie close to the diagonal of the table. The cumulative square overlap P_j is given at the bottom of Table 4. Because P_j is consistently high with values of >93%, the MBH normal modes indeed represent adequate coordinates for the motions of the lower-frequency spectrum.

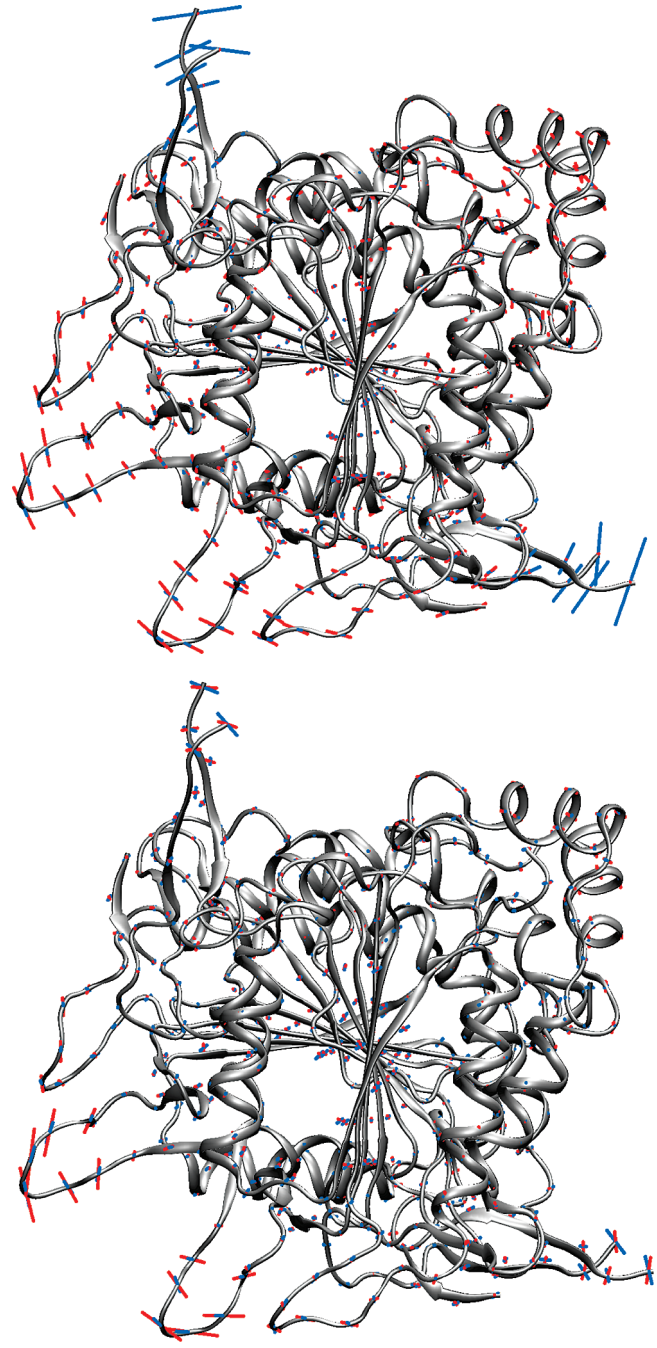


Figure 11. Caspase-1: Visualization of C_α movement in the MBH (red) and full (blue) normal modes. (top) First nonzero mode. (bottom) The fourth nonzero mode clearly shows movement of the caspase-1 subdomains with respect to each other.

The first nonzero full Hessian mode is not so well described by the lowest MBH modes. As illustrated in Figure 11, this mode is mainly active in sections where the C-terminal and N-terminal regions of two polypeptide chains meet. Other modes, on the other hand, are much better reproduced and describe more coherent, global motions of the subdomains with respect to each other.

IV. Conclusions

The extension of the MBH method to the case of adjoined blocks is a computationally attractive method for analyzing modes from the lower-frequency spectrum. The computational profit offered by the MBH with adjoined blocks can be exploited at two levels. First, during the energy minimization, fewer coordinates need to be optimized. Second, during the vibrational analysis, one needs to calculate fewer second derivatives, and the much smaller size of the adapted Hessian speeds up the diagonalization.

Testing the adjoined MBH model on several small test systems revealed that frequencies can be somewhat overestimated because of the stiffening of the system, but in general, the low-frequency normal modes are very well reproduced. The block choice is obviously crucial for the quality of the MBH frequencies and modes. By varying the block sizes, the MBH can therefore be used as an analysis tool for characterizing modes. By excluding several atoms from the blocks, MBH can also analyze the coupling between those atoms (e.g., the reactive site) and the surrounding blocks (the environment). The main application of the MBH model with adjoined blocks, however, is in the calculation of the lower spectrum of macromolecules. Especially for proteins, the MBH approach provides an efficient method to focus on the modes involved in conformational changes. Several block choices were examined for their ability to reproduce the low frequencies and modes. Among the adjoined block schemes, the $[C_\alpha]^{3b}$ block choice, where peptide units and side chains are grouped into blocks with C_α as the adjoining atom, performs well in reproducing the modes of relevance while reducing the number of modes considerably. In the $[C_\alpha{-}N]^2 + [C_\alpha{-}C]^2$ block choice, where only dihedral angles are retained as degrees of freedom, the reduction of the dimensionality is very large. The resulting modes are less accurate, but still reasonable. The MBH method with adjoined blocks therefore appears to be an appropriate analysis tool for the lower spectrum of peptide chains.

Acknowledgment. A.G. is Aspirant of the Fund for Scientific Research - Flanders (FWO). This work was supported by the Fund for Scientific Research - Flanders (FWO), the Research Board of Ghent University (BOF), and the Belgian Program on Interuniversity Attraction Poles (IAP).

References

(1) Xiao, M.; Reifenberger, J. G.; Wells, A. L.; Baldacchino, C.; Chen, L. Q.; Ge, P. H.; Sweeney, H. L.; Selvin, P. R. *Nat. Struct. Biol.* **2003**, *10* (5), 402–408.

(2) Ishima, R.; Torchia, D. A. *Nat. Struct. Biol.* **2000**, *7* (9), 740–743.

(3) Saibil, H. R. *Nat. Struct. Biol.* **2000**, *7* (9), 711–714.

(4) Rossmann, M. G.; Morais, M. C.; Leiman, P. G.; Zhang, W. *Structure* **2005**, *13* (3), 355–362.

(5) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15* (2), 151–156.

(6) Schlick, T.; Barth, E.; Mandziuk, M. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 181–222.

(7) Case, D. A. *Curr. Opin. Struct. Biol.* **1994**, *4* (2), 285–290.

(8) Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77* (9), 1905–1908.

(9) Tama, F. *Protein Pept. Lett.* **2003**, *10* (2), 119–132.

(10) Durand, P.; Trinquier, G.; Sanejouand, Y. H. *Biopolymers* **1994**, *34* (6), 759–771.

(11) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. *Proteins: Struct. Funct. Genet.* **2000**, *41* (1), 1–7.

(12) Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Verstraelen, T.; Waroquier, M. *J. Chem. Phys.* **2007**, *126* (22), 224102.

(13) Ghysels, A.; Van Neck, D.; Waroquier, M. *J. Chem. Phys.* **2007**, *127* (16), 164108.

(14) Ghysels, A.; Van Speybroeck, V.; Verstraelen, T.; Van Neck, D.; Waroquier, M. *J. Chem. Theory Comput.* **2008**, *4* (4), 614–625.

(15) Li, H.; Jensen, J. H. *Theor. Chem. Acc.* **2002**, *107*, 211–219.

(16) Besley, N. A.; Metcalf, K. A. *J. Chem. Phys.* **2007**, *126* (3), 035101.

(17) Calvin, M. D.; Head, J. D.; Jin, S. Q. *Surf. Sci.* **1996**, *345* (1−2), 161–172.

(18) Head, J. D. *Int. J. Quantum Chem.* **1997**, *65* (5), 827–838.

(19) Head, J. D.; Shi, Y. *Int. J. Quantum Chem.* **1999**, *75* (4−5), 815–820.

(20) Head, J. D. *Int. J. Quantum Chem.* **2000**, *77* (1), 350–357.

(21) Heuts, J. P. A.; Gilbert, R. G.; Radom, L. *J. Phys. Chem.* **1996**, *100* (49), 18997–19006.

(22) Van Speybroeck, V.; Van Neck, D.; Waroquier, M. *J. Phys. Chem. A* **2000**, *104* (46), 10939–10950.

(23) Van Speybroeck, V.; Van Neck, D.; Waroquier, M. *J. Phys. Chem. A* **2002**, *106* (38), 8945–8950.

(24) Van Speybroeck, V.; Vansteenkiste, P.; Van Neck, D.; Waroquier, M. *Chem. Phys. Lett.* **2005**, *402* (4−6), 479–484.

(25) Lynch, V. A.; Mielke, S. L.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109* (44), 10092–10099.

(26) Vansteenkiste, P.; Van Neck, D.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Phys.* **2006**, *124* (4), 044314.

(27) Pfaendtner, J.; Yu, X.; Broadbelt, L. J. *Theor. Chem. Acc.* **2007**, *118*, 881–898.

(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4* (2), 187–217.

(29) Vangunsteren, W. F.; Berendsen, H. J. C. *Mol. Phys.* **1977**, *34* (5), 1311–1327.

(30) Pickett, H. M.; Strauss, H. L. *J. Am. Chem. Soc.* **1970**, *92* (25), 7281.

(31) Feig, M. *J. Chem. Theory Comput.* **2008**, *4* (9), 1555−1564.

(32) Smith, P. E. *J. Chem. Phys.* **1999**, *111* (12), 5568–5579.

(33) Kwac, K.; Lee, K. K.; Han, J. B.; Oh, K. I.; Cho, M. *J. Chem. Phys.* **2008**, *128* (10), 105106.

(34) Feig, M. *J. Chem. Theory Comput.* **2007**, *3* (5), 1734–1748.

(35) Chekmarev, D. S.; Ishida, T.; Levy, R. M. *J. Phys. Chem. B* **2004**, *108* (50), 19487–19495.

(36) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward,

T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108* (21), 6582–6594.

(37) Kim, Y. S.; Wang, J. P.; Hochstrasser, R. M. *J. Phys. Chem. B* **2005**, *109* (15), 7511–7521.

(38) Grdadolnik, J.; Grdadolnik, S. G.; Avbelj, F. *J. Phys. Chem. B* **2008**, *112* (9), 2712–2718.

(39) Dadarlat, V. M. *Biophys. J.* **2005**, *89* (3), 1433–1445.

(40) Solov'yov, I. A.; Yakubovich, A. V.; Solov'yov, A. V.; Greiner, W. *Phys. Rev. E* **2006**, *73* (2), 021916.

(41) Bonvin, A.; Boelens, R.; Kaptein, R. *Biopolymers* **1994**, *34* (1), 39–50.

(42) Wilson, K. P.; Black, J. A. F.; Thomson, J. A.; Kim, E. E.; Griffith, J. P.; Navia, M. A.; Murcko, M. A.; Chambers, S. P.; Aldape, R. A.; Raybuck, S. A.; Livingston, D. J. *Nature* **1994**, *370* (6487), 270–275.

(43) Chowdhury, I.; Tharakan, B.; Bhat, G. K. *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.* **2008**, *151* (1), 10–27.

(44) Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Brooks, B. R.; Waroquier, M. *J. Chem. Phys.* **2009**, *130*, 084107.

CT800489R

JCTC Journal of Chemical Theory and Computation

Neutral and Anionic Gold Decamers: Planar Structure with Unusual Spatial Charge-Spin Separation

Young Cheol Choi,[†] Woo Youn Kim,[†] Han Myoung Lee,*[,‡] and Kwang S. Kim*[,†]

Center for Superfunctional Materials, Department of Chemistry, and Center for Basic Sciences, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

Received August 1, 2008

Abstract: We have investigated the issue of two-dimensional (2D) versus three-dimensional (3D) structures for neutral-state Au_{10} and clarified the lowest-energy structure among a few 2D Au_{10}^- isomers. Though almost all previous works were based on density functional theory (DFT), we here carried out not only extensive DFT calculations but also high levels of ab initio calculations of Möller-Plesset second order perturbation theory (MP2), and coupled cluster theory with single and double excitations (CCSD) including perturbative triple excitations [CCSD(T)]. While DFT favors 2D structures, MP2 and CCSD(T) favor 3D structures for moderate-sized basis sets. However, we note that the basis-set superposition error (BSSE) corrections make the ab intio results favor 2D structures too. The near-degeneracy (driven by relativistic effects) of 5d and 6s orbitals of gold helps stabilize acute apex gold atoms, resulting in 2D structures. The planar triangular structures of a local minimum Au_{10} (triplet) and the global minimum Au_{10}^- show remarkable spatial charge-spin separation due to their singly occupied molecular orbital(s). By the same reason, Au_{10}^- shows much larger vertical detachment energy than other even-numbered gold cluster anions.

I. Introduction

Nanoscale metal clusters[1] have received great attention as one of the most fundamental subjects of nanosize metallic materials[2] and devices.[3] These metal clusters can form an intriguing arrangement of atoms and often show unusual features for a certain specific size. Since gold is known to be a special novel metal showing effective s−d hybridization due to the large relativistic effect,[4] gold nanoclusters (Au_n) have been extensively investigated.[5,6] In particular, the lowest-energy structures of Au_{10} and Au_{10}^- are not clear. In particular, various two-dimensional (2D)[7−9] and three-dimensional (3D)[10] structures have been reported as possible global minimum energy structures for the neutral state, Au_{10}. In this regard, Au_{10} is likely to have the structures corresponding to the crossover region of the 2D to 3D transition. For the anion state, Au_{10}^-, Kappes, Ahlrichs, and co-workers[11] proved convincingly that Au_{10}^- is planar by comparing the experimentally measured ion mobility with DFT results. In addition, the presence of multiple isomers of planar Au_{10}^- is also known.[12] Landman and co-workers[8] have also reported 2D structures for Au_{10}^-. However, all these results were based on DFT calculations. Since DFT often tends to favor 2D structures over 3D structures,[7,13] a clear demonstration with ab initio methods such as Möller-Plesset second-order perturbation theory (MP2) and coupled cluster theory with single and double excitations including perturbative triple excitations [CCSD(T)] would be desirable.

In addition, the neutral and anionic gold decamers (Au_{10} and Au_{10}^-) show particularly unusual features such as large vertical-electron detachment energies (VDEs).[7−10,14,15] In general, odd-numbered gold cluster anions have large VDEs, and even-numbered gold cluster anions have small VDEs.[14,15] However, in the case of Au_{10}^-, an even-numbered cluster, the photoelectron spectrum (PES) shows a large VDE with

* Corresponding author e-mail: kim@postech.ac.kr (K.S.K.) or abcd0lhm@postech.ac.kr (H.M.L.).

[†] Center for Superfunctional Materials and Departments of Chemistry and Physics.

[‡] Department of Chemistry and Center for Basic Sciences.

10.1021/ct8003113 CCC: $40.75
Published on Web 04/24/2009

very small energy gap between the first and second main peaks.[14,15] Furthermore, we find unexpectedly large spatial charge−spin separation (SCSS) phenomena for Au_{10}^- and Au_{10}, which could be of importance for the study of spin dot devices. Here, we discuss these intriguing features, using DFT and high-level ab initio calculations.

II. Computational Details

We studied the system with DFT using two different functionals: (i) Becke's three-parameter exchange and Lee−Yang−Parr correlation functionals (B3LYP) was used with the Hay−Wadt valence double-ζ (VDZ) ($n + 1$) effective core potential (ECP) basis set (to be abbreviated as HW),[16] and (ii) the generalized gradient approximation (GGA) of Becke and Perdew for exchange and correlation functional (BPW91) was employed with the Martin (8s7p6d2f)/[6s5p3d2f] basis set[17]using the relativistic 19-electron Stuttgart−Dresden−Bonn (SDB) pseudopotentials.[18] The one-electron Darwin and mass-velocity relativistic corrections were included in this ECP. Frequency calculations and thermal energy corrections were performed at the same level of theory. The PES spectra were obtained by shifting the energy levels of occupied orbitals so that the highest occupied molecular orbital (HOMO) level was fit for VDE at a given calculation level.

To test the reliability of the calculation methods, we compared the calculated electronic properties of Au and Au_2 with experimental values. The electron affinity (EA) and ionization potential (IP) of a single Au atom at the B3LYP/HW [BPW91/HW] level are 2.17 [2.26] and 9.42 [9.50] eV, and those at the B3LYP/SDB [BPW91/SDB] level are 2.21 [2.25] and 9.43 [9.56] eV, respectively, These are in reasonable agreement with the experimental EA and IP (2.3 and 9.2 eV).[19] At the B3LYP/HW [BPW91/HW] level, the bond length and dissociation energy of Au_2 are estimated to be 2.57 [2.55] Å and 1.86 [2.03] eV, and at the B3LYP/SDB [BPW91/SDB] level those of Au_2 are 2.56 [2.54] Å and 1.92 [2.11] eV. These results are also in reasonable agreement with the experimental results, where the bond length and dissociation energy (D_0) of Au_2 are known to be 2.47 Å and 2.29 eV, respectively.[19]

We also employed MP2, CCSD, and CCSD(T) calculations. Both small HW and large SDB basis sets were used. All the MP2, CCSD, and CCSD(T) calculations were performed on the MP2/HW geometries. The BPW91/SDB results are similar to the B3LYP/HW results, and the MP2/SDB results are also similar to the MP2/HW results. Zero-point energy (ZPE) and thermal energy corrections were performed to obtain the ZPE-corrected interaction energy (ΔE_0) and free energy at 298 K and 1 atm (ΔG_{298}) at the B3LYP/HW and BPW91/SDB levels. Since the DFT results are found to be more consistent with experiment and are also closer to the CCSD and CCSD(T) results than the MP2 results, we discuss the DFT results unless otherwise specified.

However, we note that the MP2 and CCSD, CCSD(T) binding energies strongly depend on the basis set size. This indicates that even though the SDB basis set is not small, it still has large BSSEs. Thus, we carried out BSSEs using the counterpoise correction,[20] with which the total energies for 10 different structures composed of one real atom and the remaining nine dummy atoms were calculated, and the sum of them were subtracted by 10 times the energy of an isolated gold atom. In addition, to remove the finite size effect of basis sets used here, we estimated the complete basis set (CBS) limit value using the extrapolation (or partly interpolation) scheme that we previously reported.[21]

To demonstrate that the BSSE correction and the CBS extrapolation are meaningful not only for intermolecular interactions but also for strongly bound systems like the gold clusters, we calculated the binding energy of the gold dimer (Au_2) at the CCSD level using the HW and SDB basis sets, where the geometry was optimized at the CCSD/SDB level. The calculated bond length is 2.52 Å, comparable to the experimental value of 2.47 Å.[22] The binding energies are 1.80 and 1.96 eV for the HW and SDB basis sets, respectively, while the experimental value is 2.29 eV.[22] The estimated CBS limit value is 2.46 eV, which was closer to the experimental value. This demonstrates that the CBS estimation is reliable, and it is thus applicable to strongly bound systems like gold clusters. The extrapolation employed here is practically or to some extent an interpolation scheme as compared with other methods, which are solely extrapolations that do not give any proper bound values. It is true that the HW and SDB basis sets themselves give rather poor binding energies. In this regard, the CBS limit value is one of the alternatives to get more reliable binding energies for large systems like the gold decamer.

All the DFT and MP2 energy calculations were performed with the Gaussian suite of programs,[23] and all the CCSD and CCSD(T) calculations were performed with the MOLPRO package.[24] Charge and spin density were calculated with both the Gaussian[23] and VASP[25] programs. B3LYP/HW and BPW91/SDB calculations were carried out with atomic orbital basis sets. The plane wave calculations were carried out with the ultrasoft Vanderbilt pseudopotential and PW91 GGA exchange−correlation functional. Cluster structures were drawn with the Posmol package.[26]

III. Results and Discussion

A. Structures. To search for the lowest-energy structures for the neutral and anionic states of the gold decamer, we have carried out calculations for their various structures of different topologies at varied levels of theory. The low-lying energy structures of Au_{10} and Au_{10}^- are shown in Figure 1 and their relative energies, EAs, IPs, and VDEs are listed in Table 1, where we report only two competing 3D structures (3Da and 3Db) in the neutral state and three competing 3D structures (3Db, 3Dc, and 3Dd) in the anionic state, including the lowest-energy structures among several 3D structures at the MP2/SDB and BPW91/SDB levels.

For the low-lying 2D energy structures of Au_{10} with D_{3h}(triplet)_[111], C_{2v}, C_s, and D_{2h}_[100] symmetries (Figure 1a), where the crystal projection planes are given in brackets, the B3LYP/HW [BPW91/SDB] relative energies are 0.93 [0.82], 0.10 [0.57], 0.08 [0.19], and 0.00 [0.00] eV at 0 K, respectively, and 0.91 [0.79], 0.15 [0.62], 0.00 [0.11], and 0.01 [0.00] eV at 298 K, respectively. Thus, the temperature

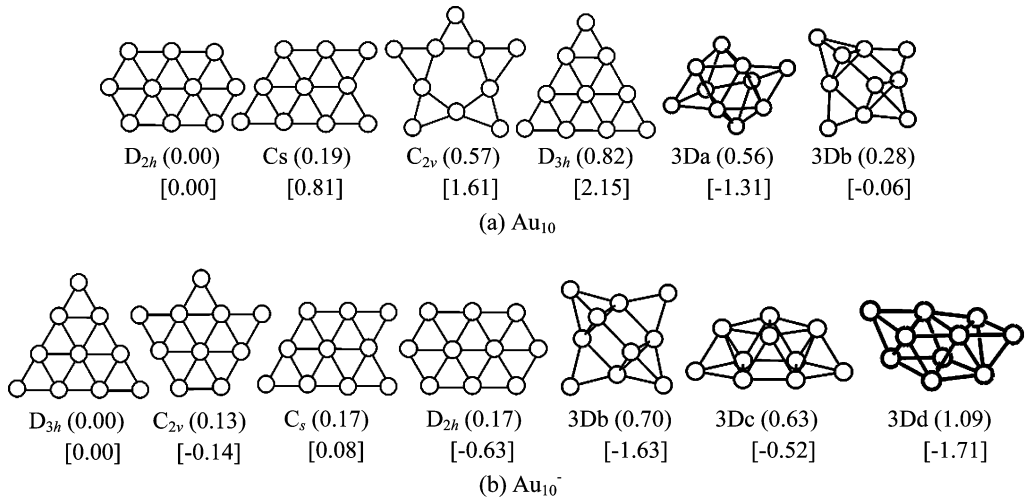


Figure 1. Various gold cluster structures with different topologies of (a) Au_{10} and (b) Au_{10}^-. BPW91/SDB [MP2/SDB] relative energies are in electronvolts in parentheses [brackets]. The given 3D structures are of the lowest energies among ~20 different 3D structures investigated here.

effect is very small, and the stability will be simply discussed based on 0 K. The corresponding MP2/HW [MP2/SDB] relative energies are 1.81 [2.15], 1.15 [1.61], 0.69 [0.81], and 0.00 [0.00] eV at 0 K, respectively. DFT and MP2 calculations are consistent, but the MP2 calculations stabilize the D_{2h} structure more.

As for Au_{10}^-, the B3LYP/HW [BPW91/SDB] relative energies of the D_{3h}_(triplet), C_{2v}, C_s, and D_{2h} structures (Figure 1b) are 0.00 [0.00], 0.04 [0.13], 0.08 [0.17], and 0.24 [0.17] eV at 0 K, respectively, and 0.00 [0.00], 0.03 [0.12], 0.07 [0.18], and 0.27 [0.20] eV at 298 K, respectively. Thus, the temperature effect is again very small. The corresponding MP2/HW [MP2/SDB] relative energies are 0.00 [0.00], −0.10 [−0.14], 0.04 [0.08], and −0.55 [−0.63] eV, respectively. Thus, in the case of Au_{10}^-, the ordering of relative energies for the above four clusters is completely reversed between the DFT and MP2 results.

On metal clusters like Au_8, we note that, as compared with the MP2 results, the DFT results are more consistent with very accurate calculations such as CCSD(T), while the CCSD results are almost in between the DFT and MP2 results.[27] Nevertheless, in consideration that DFT often fails in properly describing anionic clusters, we have carried out CCSD//MP2 calculations using both the HW and SDB ECPs. The CCSD results for Au_{10}^- show that the D_{3h} structure is 0.04 eV more stable than the D_{2h} structure, which is more consistent with the DFT results. Therefore, even for the anionic gold clusters, the DFT results are more reliable than the MP2 results. This will be further verified by the electronic properties and PES spectra, which will be discussed later.

At the B3LYP/HW [BPW91/SDB] level, the 2D structures are favored for both neutral and anionic states; the 3D lowest-energy structure is 0.09/0.53 [0.28/0.63] eV higher than the 2D lowest-energy structure in the neutral/anionic state. However, at the MP2/HW [MP2/SDB] level, the 3D structures are favored for both neutral and anionic states; the 3D lowest-energy neutral/anionic structure is 1.11/0.89 [1.31/1.08] eV more stable than the 2D lowest-energy neutral(D_{2h})/anionic(D_{3h}) structure. It is well-known that, in cluster calculations, MP2 results tend to favor 3D structures over 2D structures, while DFT results are often in favor of 2D structures.[28] On the other hand, in the CCSD/HW [CCSD/SDB] calculations of the neutral state, the 3D structure is [0.70/0.30] eV more stable than the 2D D_{2h} structure. In the CCSD/HW calculations of the anionic state, the 3D structure is 0.58 eV more stable than the 2D D_{3h} structure, whereas the 2D D_{3h} structure is 0.97 eV more stable than the 3D structure in the CCSD/SDB calculations. For the CCSD(T)/HW calculations of the neutral/anionic state, the 3D structure is 0.91 and 0.87 eV more stable than the 2D D_{2h}/D_{3h} structure.

However, the 3D structures tend to have larger BSSEs than 2D structures, so we have done the BSSE correction for the 2D and 3D lowest-energy structures at the CCSD and CCSD(T) levels. The stabilities for 2D and 3D structures of neutral and anionic states are reversed after the BSSE correction except for the structures of anionic states at the CCSD/SDB level. The 2D lowest-energy neutral(D_{2h})/anionic(D_{3h}) structure is 0.18/0.32 [0.01/0.07] eV more stable than the 3D lowest-energy structure at the CCSD/HW [CCSD(T)/HW] level and 0.31/1.24 eV more stable than the 3D structure at the CCSD/SDB level.

We also estimated the CBS limit value. In the case of neutral state, the BSSE corrected and uncorrected energies for the two different HW and SDB basis sets were used to estimate the CCSD/CBS value. Then, the neutral/anionic 2D structure is 0.60/1.63 eV more stable than the 3D lowest-energy structure. The BSSE uncorrected [corrected] relative energy of the anionic 2D D_{3h} structure with respect to the 3D lowest-energy structure (3Dc) is 0.58 [−0.32] eV, whereas that of the neutral 2D D_{2h} structure is 0.70 [−0.18] eV.

Thus, the BSSE correction enhances the stability of anionic/neutral 2D structure over the 3D structure by 0.90/0.88 eV, which indicates that the BSSE correction effect is almost the same for both anionic and neutral systems. Since the BSSEs for MP2, CCSD, and CCSD(T) calculations are similar in most cases, in the neutral [anionic] state the relative

Table 1. Relative Binding Energies, Vertical Electron Affinities, Vertical Ionization Potentials, and Vertical Detachment Energies of Au_{10} and Au_{10}^- [a]

Au_{10}: ΔE_0 (ΔG_{298})	$D_{3h}^{(t)}$	C_{2v}	C_s	D_{2h}	3Da	3Db
B3LYP/HW	0.93 (0.91)	0.10 (0.15)	0.08 (0.00)	0.00 (0.01)	0.59 (0.60)	0.09 (0.05)
BPW91/SDB	0.82 (0.79)	0.57 (0.62)	0.19 (0.11)	0.00 (0.00)	0.56 (0.58)	0.28 (0.24)
MP2/HW	1.81	1.15	0.69	0.00	−1.11	−0.08
MP2/SDB	2.15	1.61	0.81	0.00	−1.31	−0.06
CCSD/HW				0.00	−0.70	−0.31
CCSD/HW-BSSE				0.00	0.18	0.28
CCSD/SDB				0.00	−0.30	
CCSD/SDB-BSSE				0.00	0.31	
CCSD/CBS				0.00	0.60	
CCSD(T)/HW				0.00	−0.91	−0.27
CCSD(T)/HW-BSSE				0.00	0.01	0.34
CCSD(T)/CBS				0.00	0.43	

Au_{10}^-: ΔE_0 (ΔG_{298})	D_{3h}	C_{2v}	C_s	D_{2h}	3Db	3Dc	3Dd
B3LYP/HW	0.00 (0.00)	0.04 (0.03)	0.08 (0.07)	0.24 (0.27)	0.53 (0.52)	0.75 (0.76)	1.10 (1.07)
BPW91/SDB	0.00 (0.00)	0.13 (0.12)	0.17 (0.18)	0.17 (0.20)	0.63 (0.62)	0.70 (0.71)	1.09 (1.07)
MP2/HW	0.00	−0.10	0.04	−0.55	−0.39	−1.35	−1.44
MP2/SDB	0.00	−0.14	0.08	−0.63	−0.52	−1.63	−1.71
CCSD/HW	0.00			0.04	−0.08	−0.58	0.27
CCSD/HW-BSSE	0.00					0.32	
CCSD/SDB	0.00					0.97	
CCSD/SDB-BSSE	0.00					1.24	
CCSD/CBS	0.00					1.63	
CCSD(T)/HW	0.00				−0.17	−0.87	
CCSD(T)/HW-BSSE	0.00					0.07	
CCSD(T)/CBS	0.00					1.38	

Au_{10}: EA_v	$D_{3h}^{(t)}$	C_{2v}	C_s	D_{2h}	3Da	3Db
B3LYP/HW	4.01	3.09	3.06	2.82	2.47	2.54
BPW91/SDB	3.75	3.19	2.72	2.76		

Au_{10}: IP_v	$D_{3h}^{(t)}$	C_{2v}	C_s	D_{2h}	3Da	3Db
B3LYP/HW	7.59	8.07	7.50	7.70	7.26	7.11
BPW91/SDB	7.51	7.42	7.31	7.48	7.23	7.04
MP2/HW	6.88	7.23	7.02	7.52	6.92	6.61
MP2/SDB	7.58	8.66	7.98	8.15	7.79	7.63
CCSD/HW				7.05	6.61	6.38

Au_{10}^-: VDE	D_{3h}	C_{2v}	C_s	D_{2h}	3Db	3Dc	3Dd
B3LYP/HW	4.17	3.55	3.26	3.02	3.22	2.90	2.83
BPW91/SDB	3.98	3.50	3.13	2.93	3.06	2.84	2.62
MP2/HW	3.76		2.59	2.49	2.59	2.36	2.58
MP2/SDB	4.24	3.78	3.23	3.08	3.20	3.02	3.18
CCSD/HW	3.53				2.63	2.27	1.74
CCSD(T)/HW	3.64				2.72	2.44	

[a] Values in parentheses are the relative free energies at room temperature and 1 atm (ΔG_{298}); otherwise, values denote the zero-point-energy-corrected relative energy (ΔE_0). The experimental IP of Au_{10} is 8.2 eV and photoelectron spectra show small peaks around 2.9−3.5 eV and main peaks around 4.0 eV.[14,15] Superscript (t) indicates the triplet spin state, and suffix -BSSE indicates the BSSE-corrected relative energy. See text for estimation of the CBS energy.

energy of the 3Da [3Dc] structure at the CCSD(T)/HW level is likely to be 0.17 [0.25] eV lower than that at the CCSD/HW level. Thus, at the CCSD(T)/CBS level, the neutral and anionic 2D structures are 0.43 and 1.38 eV more stable than the corresponding 3D structures, respectively. This indicates that both neutral and anionic Au_{10} have the 2D structures, which will be further verified by comparing the predicted IP and VDE with the experimental data.

B. Ionization Potential and Vertical Detachment Energy. It is noted that the IPs are underestimated at the B3LYP/HW, BPW91/SDB, and MP2 levels. In the case of Au_8,[7] the experimental IP is 8.65 eV, while the predicted IP at the B3LYP/HW [BPW91/SDB] level is 8.11 [8.03] eV, which is underestimated by 0.54 [0.62] eV. Thus, the predicted IP at this level of theory is underestimated. On the other hand, the predicted VDEs are in good agreement with experiments. For Au_8^-,[8] the predicted VDE is 2.94 eV, in good agreement with the experimental VDE of 2.79 eV. In this regard, the VDE is the useful information to find the right structure.

The IPs of 2D D_{2h} and 3Da structures of Au_{10} are estimated to be 7.70 [7.48, 7.49] and 7.26 [7.23, 7.07] eV at the B3LYP/HW [BPW91/SDB, B3LYP/SDB] level and 7.49 [7.24] eV at the B3LYP/SDB level, while the experimental

IP of Au_{10} is known to be 8.2 eV.[14,15] Thus, in the case of 3D structure, the predicted IPs are underestimated by 0.94 [0.97] eV at the B3LYP/HW [BPW91/SDB] level, whereas in the case of 2D structure, the predicted IPs are underestimated by 0.50 [0.72] eV at the B3LYP/HW [BPW91/SDB] level. In addition, the IPs of 2D D_{2h} and 3Da structures are 7.52 [8.15] and 6.92 [7.79] eV at the MP2/HW [MP2/SDB] level, respectively. Thus, the MP2/SDB IP of 2D D_{2h} (8.15 eV) agrees well with the experimental value (8.2 eV). Nevertheless, the corresponding values for CCSD(HW) are 7.05 and 6.61 eV, in poor agreement, possibly due to the insufficient basis set at the CCSD level. As the long-range dispersion effect such as aurophilic effects[29] would be important in the gold systems, the DFT IPs would be underestimated. Nevertheless, it is not clear only from the present data because we were not able to obtain the IPs at the CCSD(T) level using a large basis set. The available data imply that the IPs for 2D structure lie in a more probable range than the 3D structure. This could indicate that the lowest-energy structure of neutral Au_{10} is likely to be 2D D_{2h}. On the other hand, the calculated B3LYP/HW [BPW91/SDB] PES spectra of Au_{10}^- isomers match the experimental[8,14,15] ones. The experimental PES spectrum of Au_{10}^- shows small peaks from 2.9 to 3.5 eV, which correspond to the predicted isomers of C_{2v} (3.55 [3.50] eV), C_s (3.26 [3.13] eV), and D_{2h} (3.02 [2.93] eV) structures, while the main peaks around 4.0 eV correspond to the minimum-energy isomer having the D_{3h} (4.17 [3.98] eV) structure. Moreover, the relative intensities of the experimental PES peaks, which reflect the populations of the given clusters, indicate that the 2D structure with D_{3h} symmetry would be the lowest-energy structure of Au_{10}^-. None of the low-energy 3D anionic structures matches the experimental VDE.

C. Origin of Stability and Large Vertical Detachment Energy. To understand why the 2D structures are more stable than the 3D structures in both neutral and anionic states, we investigate the same planar D_{2h} structures of Au_{10}_D_{2h}, Ag_{10}_D_{2h}, and Na_{10}_D_{2h} by carrying out the BPW91/SDB calculations for Au_{10} and Ag_{10}, and the BPW91/cc-pCVDZ calculations for Na_{10} (as the SDB basis set is not available for Na). The total charges on the six apex atoms of Au_{10}_D_{2h}, Ag_{10}_D_{2h}, and Na_{10}_D_{2h} are −0.21, −0.15, and −0.15, respectively. Their charge populations of s/p/d orbitals of the HOMO state are 0.74/0.02/0.24, 0.77/0.17/0.06, and 0.79/0.21/0.00, respectively. The apex atoms of Au_{10}_D_{2h} have a significant charge population of d orbitals in contrast to those of Na_{10}_D_{2h}, which have a significant population of p orbitals. For Au_{10}_D_{2h}, the charge repulsions become minimal because the distances between different apexes are much more separated than the distances between any other sites. This demonstrates that the near-degeneracy (driven by relativistic effects) of 6s and 5d orbitals plays an important role in stabilizing 2D Au clusters, which is almost absent in Na and Ag. We also investigated the same planar triangular structures of Au_{10}^-_D_{3h}, Ag_{10}^-_D_{3h}, and Na_{10}^-_D_{3h}. The charges of an apex atom for Au_{10}^-_D_{3h}, Ag_{10}^-_D_{3h}, and Na_{10}^-_D_{3h} are −0.29, −0.25, and −0.25, respectively. Their charge populations of s/p/d orbitals of the singly occupied molecular orbital (SOMO) states are 0.68/0.02/0.30, 0.75/0.18/0.07, and 0.82/0.18/0.00, respectively, which also shows the near-degeneracy (driven by relativistic effects) of 6s and 5d orbitals. In Au_{10}^-_D_{3h}, the excess electron is also mostly populated on the apex atoms in the triangular structure. Since the Coulombic repulsions between apex atoms (which are more separated than any other two atoms) are weaker, this energy gain makes the 2D D_{3h} structure more stabilized than other structures, which would result in lowering the SOMO energy and hence increasing VDE.

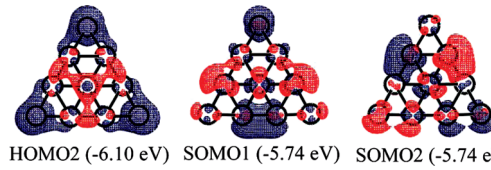


Figure 2. Frontier molecular orbitals of Au_{10}_D_{3h} at the BPW91/SDB level. The orbital energy is given in parentheses.

Figure 2 shows the molecular orbital of Au_{10} at the BPW91/SDB level. Au_{10}^- is predicted to have a small energy separation (less than 0.1 eV) between the first and second PES peaks, in agreement with the experimental PES of Au_{10}^-.[8,14] This reflects why the ground-state neutral structure (D_{3h}) of Au_{10}, which has the same triangular geometry with the lowest-energy structure (D_{3h}) of anionic Au_{10}^-, is a triplet. A large vertical electron affinity (EA_v) of the neutral structure Au_{10}_D_{3h} is due to two degenerate low-energy SOMOs. Since an additional electron occupies one of the two low-energy SOMOs, Au_{10}^- has a large VDE in contrast to the case when the electron occupies the lowest unoccupied molecular orbital (LUMO). Thus, the VDE of Au_{10}^- is large, unlike other ordinary even-numbered gold clusters whose VDEs are small. Though Au_{10}^- maintains a triangular structure, its symmetry changes from D_{3h} to C_{2v}^{JT} (though very close to D_{3h}) for better stabilization due to the Jahn−Teller (JT) distortion.

D. Charge−Spin Separation. Natural bond orbital (NBO) charges (Q) and spin populations (S) in Au_{10} and Au_{10}^- calculated at the BPW91/SDB level are in Table 2. An excess electron can go into one of the doubly degenerate SOMO levels of Au_{10}_D_{3h} with equal probabilities. The charge of an edge atom is either positive (i.e., electron-deficient; 0.07 for Au_{10}_D_{3h}) or nearly zero (−0.01 for Au_{10}^-_C_{2v}^{JT}), while the charge of an apex atom is highly negative (−0.13 au for Au_{10}_D_{3h} and −0.29 au for Au_{10}^-_C_{2v}^{JT}). Charge density is separated into regions with either positively or minimally negatively charged edges and regions with highly negatively charged apexes, which results in maximizing the electrostatic energy gain.

Here, we note a fascinating phenomenon in Au_{10}_D_{3h} and Au_{10}^-_C_{2v}^{JT}, whose spin density is small around the central and apex atoms but high around edge atoms, as shown in Table 2. Thus, the spin density is populated almost oppositely to the charge density. To verify these results, we also performed charge and spin density calculations using the plane wave pseudopotential methods (instead of the Gaussian type orbital methods) with generalized gradient approximation to exchange−correlation using the VASP package. These charge and spin densities (Figure 3) are consistent with those based on BPW91/SDB with the Gaussian orbitals (Table 2).

Table 2. Charge and Spin of Each Type of Atom for Six Triangular Structures[a]

	$Au_{10}^-\_C_{2v}^{JT}$		$Au_{10}^{(t)}\_D_{3h}$		$Na_{10}^{(t)}\_D_{3h}$		$Ag_{10}^{(t)}\_D_{3h}$		$Pt_{10}^{(s)}\_C_{2v}$		$Au_{15}^{(d)}\_C_{2v}$	
atoms	*Q*	*S*	*Q*	*S*	*Q*	*S*	*Q*	*S*	*Q*	*S*	*Q*	*S*
center	0.01	0.11	−0.03	0.16	−0.25	−0.10	−0.19	−0.02	−0.12	0.18	0.01	0.05
edge	−0.02	0.12	0.07	0.24	0.07	0.25	0.07	0.23	0.09	0.58	0.04	0.04
apex	−0.29	0.06	−0.13	0.14	−0.05	0.20	−0.07	0.20	−0.13	0.78	−0.14	0.16

[a] Charge (*Q*) and spin (*S*) are given in atomic units. Superscripts (t), (s), and (d) denote triplet, septet, and doublet spin states, respectively, and subscript JT denote the Jahn−Teller distortion. C_{2v}^{JT} is very close in structure to D_{3h}. $Au_{10}^-\_C_{2v}^{JT}$, $Au_{10}^{(t)}\_D_{3h}$, $Ag_{10}^{(t)}\_C_{2v}$, $Pt_{10}^{(s)}\_C_{2v}$, and $Au_{15}^{(d)}\_C_{2v}$ were calculated at the BPW91/SDB level and $Na_{10}^{(t)}\_D_{3h}$ was calculated at the BPW91/cc-pCVDZ level.

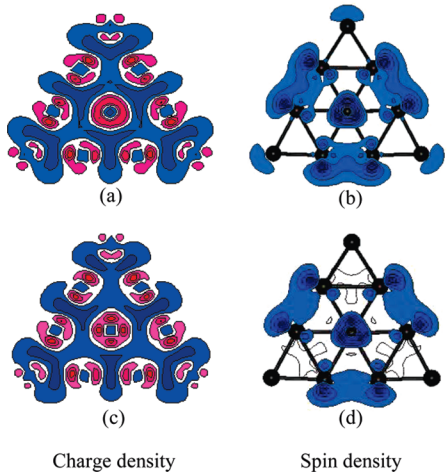


Figure 3. Charge density and spin density for (a, b) $Au_{10}\_D_{3h}$ and (c, d) $Au_{10}^-\_C_{2v}^{JT}$ (which is similar to D_{3h}). The difference of the total charge density from the sum of charge densities of all isolated gold atoms at each position of Au_{10} is plotted in contour map, based on the plane wave DFT calculations, where blue indicates regions with rich electron or spin density, while red indicates regions with depleted electron density.

The spin densities arise mainly from frontier singly occupied orbitals. For $Au_{10}\_D_{3h}$, both SOMOs (SOMO1 and SOMO2) are mainly distributed around the edge atoms, resulting in larger spin density on the edge atoms. On the other hand, the charge density is mostly populated around the apex atoms because of the doubly occupied molecular orbitals below SOMOs. Therefore, we note the substantial SCSS, which is extremely rare. When an excess electron is added to $Au_{10}\_D_{3h}$, it occupies one of the SOMOs; thus the total spin of $Au_{10}^-\_D_{3h}$ (which eventually changes to $Au_{10}^-\_C_{2v}^{JT}$ due to the JT effect) is reduced from 2 to 1, partly suppressing the spin density. Nevertheless, the SCSS is still clear in $Au_{10}^-\_D_{3h}$ and $Au_{10}^-\_C_{2v}^{JT}$.

By including spin−orbit coupling, magnetization was evaluated for the neutral-state 2D D_{3h} structure using plain wave pseudopotential density functional methods. The average magnetization per atom on the apex atoms is only 4.6% of total magnetization, while that on the edge atoms is 11.3%, and the unoccupied states are located in a very low energy region only 0.25 eV higher than the occupied states, which shows that even though the spin−orbit coupling is included, the characteristic electronic structures for the 2D D_{3h} structure are maintained.

For comparison, we investigated the planar structures of Au_{10} clusters with D_{2h}, C_s, and C_{2v} symmetries shown in Figure 1. These do not show SCSS because these lowest-energy spin states are singlet states. Planar triangular $Au_6\_D_{3h}$ and $Au_{15}\_D_{3h}$ were also investigated to verify whether planar triangular symmetry is the necessary condition for SCSS. $Au_6\_D_{3h}$ has a singlet spin state, resulting in no SCSS, while $Au_{15}\_D_{3h}$ has a doublet spin state with large densities of both charge and spin on apex atoms and does not show unusual SCSS because the charge goes with the spin (Table 2). This indicates that the degenerate triplet state of 10 metal atoms cluster with planar triangular D_{3h} symmetry results in substantial SCSS. In this regard, we studied other metal systems with planar triangular structures of $Ag_{10}\_D_{3h}$ and $Na_{10}\_D_{3h}$, which satisfy the conditions for SCSS as mentioned above. They also showed SCSS, though less than $Au_{10}\_D_{3h}$, as shown in Table 2. We also investigated planar triangular $Pt_{10}\_D_{3h}$ (or $Pt_{10}\_C_{2v}^{JT}$). This is slightly distorted to have C_{2v} symmetry, still maintaining the triangular geometry. In contrast to $Ag_{10}\_D_{3h}$ and $Na_{10}\_D_{3h}$, which show significant SCSS (though less than $Au_{10}\_D_{3h}$), $Pt_{10}\_C_{2v}^{JT}$ (which has almost the same planar triangular geometry as $Au_{10}\_D_{3h}$) has different electronic configurations and a different spin state (septet) from $Au_{10}\_D_{3h}$. It shows both large charge and spin densities on apex atoms but no significant SCSS. Though $Ag_{10}\_D_{3h}$ and $Na_{10}\_D_{3h}$ show some SCSS, these clusters can hardly be made because they are much less stable than the global minimum-energy structure, while $Au_{10}\_D_{3h}$ can be synthesized by detaching one electron from the experimentally identified lowest-energy structure $Au_{10}^-\_D_{3h}$.

As shown in $Au_{10}^-\_D_{3h}$, this kind of spin polarization is called s-electron ferromagnetism,[30] where several degenerate HOMO levels due to highly symmetric geometry are partially occupied with s electrons whose spins are aligned according to Hund's rule. Therefore, such s-electron ferromagnetism is realized in the planar $Au_{10}\_D_{3h}$ cluster, while in this case the nearly degenerate d-orbitals also play a role in enhancing the ferromagnetism. Moreover, the spin on planar $Au_{10}\_D_{3h}$ would not be seriously influenced by surrounding environments, because its spin is mainly distributed around edge atoms due to SCSS, suggesting a possible use of the $Au_{10}\_D_{3h}$ cluster as a spin quantum dot.

IV. Conclusion

For gold clusters, the near-degeneracy (driven by relativistic effects) of 5d and 6s orbitals helps stabilize acute apex gold atoms. Thus, small gold clusters tend to have planar structures, and the transition from 2D to 3D structural transformation is considered to be around decamers. For Au_{10} and Au_{10}^-, DFT (B3LYP/HW and PBW91/SDB) predicts 2D structures, while MP2 (MP2/HW and MP2/SDB) predicts 3D structures. At the CCSD/HW and CCSD(T)/HW levels

of theory, Au_{10} and Au_{10}^- are also predicted to have 3D structures; at the CCSD/SDB level, Au_{10} favors 3D, while Au_{10}^- favors 2D. However, according to the estimated CCSD(T)/CBS limit with the BSSE correction, it is predicted that both Au_{10} and Au_{10}^- favor 2D structures.

The VDEs of the 2D anionic gold decamer isomers studied here explain a few peaks in the experimental PES spectra, while the $Au_{10}^-\_D_{3h}$ VDE peak is clearly dominant. Upon binding an excess electron, the degenerate state of D_{3h} symmetry transforms Au_{10}^- into a structure of C_{2v} symmetry due to the JT distortion. The calculated VDE for $Au_{10}^-\_C_{2v}^{JT}$ (close to D_{3h} symmetry) is in good agreement with the experimental value, while the VDEs of competing structures agree with a few minor peaks in the experimental PES spectra. The unusual small separation between the first and second peaks in the PES is found to be due to the doubly degenerate low-energy SOMOs of $Au_{10}^-\_D_{3h}$. Though even-numbered gold clusters usually have small VDE, $Au_{10}^-\_D_{3h}$ shows relatively large VDE due to the same reason. $Au_{10}\_D_{3h}$ and $Au_{10}^-\_C_{2v}^{JT}$ shows s-electron ferromagnetism and remarkable spatial charge-spin separations, which could be useful in spintronics.[31]

Acknowledgment. This work was supported by the projects supported by GRL/KICOS, KOSEF (WCU, R32-2008-000-10180-0, and EPB Center, R11-2008-052-01000), BK21, and KRF-2006-353-C00022 (MOEHRD) and KISTI (KSC-2008-K08-0002 and KSC-2007-S00-3005).

References

(1) (a) Häkkinen, H.; Moseler, M.; Landman, U. *Phys. Rev. Lett.* **2002**, *89*, 033401. (b) Pyykkö, P. *Angew. Chem., Int. Ed.* **2002**, *41*, 2174. (c) Böhme, D. K.; Schwarz, H. *Angew. Chem., Int. Ed.* **2005**, *44*, 2336. (d) Wesendrup, R.; Hunt, T.; Schwerdtfeger, P. *J. Chem. Phys.* **2000**, *112*, 9356. (e) Yoon, J.; Kim, K. S.; Baeck, K. K. *J. Chem. Phys.* **2000**, *112*, 9335. (f) Zheng, J.; Petty, J. T.; Dickson, R. M. *J. Am. Chem. Soc.* **2003**, *125*, 7780. (g) Pyykkö, P. *Nat. Nanotechnol.* **2007**, *2*, 273.

(2) (a) Barnett, R. N.; Landman, U. *Nature (London)* **1997**, *387*, 788. (b) Häkkinen, H.; Barnett, R. N.; Landman, U. *J. Phys. Chem. B* **1999**, *102*, 8814. (c) Rodrigues, V.; Fuhrer, T.; Ugarte, D. *Phys. Rev. Lett.* **2000**, *85*, 4124. (d) Hong, B. H.; Bae, S. C.; Lee, C.; Jeong, S.; Kim, K. S. *Science* **2001**, *294*, 348. (e) Rodrigues, V.; Bettini, J.; Silva, P. C.; Ugarte, D. *Phys. Rev. Lett.* **2003**, *91*, 096801. (f) Geng, W. T.; Kim, K. S. *Phys. Rev. B* **2003**, *67*, 233403. (g) Nautiyal, T.; Rho, T. H.; Kim, K. S. *Phys. Rev. B* **2004**, *69*, 193404. (h) Lee, Y.-J.; Brandbyge, M.; Puska, M. J.; Taylor, J.; Stokbro, K.; Nieminen, R. M. *Phys. Rev. B* **2004**, *69*, 125409. (i) Cheng, D. Y.; Kim, W. Y.; Min, S. K.; Nautiyal, T.; Kim, K. S. *Phys. Rev. Lett.* **2006**, *96*, 096104. (j) Han, Y.-K.; Jung, J. *J. Chem. Phys.* **2006**, *125*, 084101.

(3) (a) Kim, W. Y.; Choi, Y. C.; Kim, K. S. *J. Mater. Chem.* **2008**, *18*, 4510. (b) Singh, N. J.; Lee, E. C.; Choi, Y. C.; Lee, H. M.; Kim, K. S. *Bull. Chem. Soc. Jpn.* **2007**, *80*, 1437.

(4) (a) Pyykkö, P. *Chem. Rev.* **1988**, *88*, 563. (b) Schwerdtfeger, P.; Dolg, M.; Schwarz, W. H. E.; Bowmaker, G. A.; Boyd, P. D. W. *J. Chem. Phys.* **1989**, *91*, 1762. (c) Pyykkö, P. *Angew. Chem., Int. Ed.* **2002**, *41*, 3573. (d) Nautiyal, T.; Youn, S. J.; Kim, K. S. *Phys. Rev. B* **2003**, *68*, 033407. (e) Choi, Y. C.; Lee, H. M.; Kim, W. Y.; Kwon, S. K.; Nautiyal, T.; Cheng, D. Y.; Vishwanathan, K.; Kim, K. S. *Phys. Rev. Lett.* **2007**, *98*, 076101.

(5) (a) Li, J.; Li, X.; Zhai, H.; Wang, L.-S. *Science* **2003**, *299*, 864. (b) Schwerdtfeger, P. *Angew. Chem., Int. Ed.* **2003**, *42*, 1892. (c) Bonacic-Koutecky, V.; Burda, J.; Mitric, R.; Ge, M.; Zampella, G.; Fantucci, P. *J. Chem. Phys.* **2002**, *117*, 3120. (d) Pyykkö, P. *Angew. Chem., Int. Ed.* **2004**, *43*, 4412. (e) Yoon, B.; Häkkinen, H.; Landman, U.; Worz, A. S.; Antonietti, J.-M.; Abbet, S.; Judai, K.; Heiz, U. *Science* **2005**, *307*, 403. (f) Olson, R. M.; Varganov, S.; Gordon, M. S.; Metiu, H.; Chretien, S.; Piecuch, P.; Kowalski, K.; Kucharski, S. A.; Musial, M. *J. Am. Chem. Soc.* **2005**, *127*, 1049. (g) Lee, H. M.; Diefenbach, M.; Suh, S. B.; Tarakeshwar, P.; Kim, K. S. *J. Chem. Phys.* **2005**, *123*, 743281. (h) Walker, A. V. *J. Chem. Phys.* **2005**, *122*, 943101. (i) Remacle, F.; Kryachko, E. S. *J. Chem. Phys.* **2005**, *122*, 443041. (j) Gronbeck, H.; Broqvist, P. *Phys. Rev. B* **2005**, *71*, 734081. (k) Ji, M.; Gu, X.; Li, X.; Gong, X. G.; Li, J.; Wang, L.-S. *Angew. Chem., Int. Ed.* **2005**, *44*, 7119. (l) Ricci, D.; Bongiorno, A.; Pacchioni, G.; Landman, U. *Phys. Rev. Lett.* **2006**, *97*, 036106. (m) Han, Y.-K. *J. Chem. Phys.* **2006**, *124*, 243161. (n) Bulusu, S.; Li, X.; Wang, L. S.; Zeng, X. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8326. (o) Bulusu, S.; Li, X.; Wang, L.-S.; Zeng, X. C. *J. Phys. Chem. C* **2007**, *111*, 4190.

(6) (a) Lee, Y. S.; Ermler, W. C.; Pitzer, K. S.; McLean, A. D. *J. Chem. Phys.* **1979**, *70*, 288. (b) Ermler, W. C.; Lee, Y. S.; Pitzer, K. S. *J. Chem. Phys.* **1979**, *70*, 293. (c) Ziegler, T.; Snijders, J. G.; Baerends, E. J. *Chem. Phys. Lett.* **1980**, *75*, 1. (d) Ziegler, T.; Snijders, J. G.; Baerends, E. J. *J. Chem. Phys.* **1981**, *74*, 1271. (e) Lee, H.-S.; Han, Y-K.; Kim, M. C.; Bae, C.; Lee, Y. S. *Chem. Phys. Lett.* **1998**, *293*, 97. (f) Pyykkö, P.; Zhao, Y. *Angew. Chem., Int. Ed.* **1991**, *30*, 604. (g) Li, J.; Pyykkö, P. *Chem. Phys. Lett.* **1992**, *197*, 586. (h) Häkkinen, H. *Chem. Soc. Rev.* **2008**, *37*, 1847.

(7) Lee, H. M.; Ge, M.; Sahu, B. R.; Tarakeshwar, P.; Kim, K. S. *J. Phys. Chem. B* **2003**, *107*, 9994.

(8) Häkkinen, H.; Yoon, B.; Landman, U.; Li, X.; Zhai, H.; Wang, L.-S. *J. Phys. Chem. A* **2003**, *107*, 6168.

(9) Häkkinen, H.; Abbet, S.; Sanshez, A.; Heiz, U.; Landman, U. *Angew. Chem., Int. Ed.* **2003**, *42*, 1297.

(10) (a) Wang, J.; Wang, G.; Zhao, J. *Phys. Rev. B* **2002**, *66*, 035418. (b) Häkkinen, H.; Landman, U. *Phys. Rev. B* **2000**, *62*, R2287. (c) Soulé de Bas, B.; Ford, M. J.; Cortie, M. B. *J. Mol. Struct. (THEOCHEM)* **2004**, *686*, 193.

(11) Furche, F.; Ahlrichs, R.; Weis, P.; Jacob, C.; Gilb, S.; Bierweiler, T.; Kappes, M. M. *J. Chem. Phys.* **2002**, *117*, 6982.

(12) Gilb, S.; Jacobsen, K.; Schooss, D.; Furche, F.; Ahlrichs, R.; Kappes, M. M. *J. Chem. Phys.* **2004**, *121*, 4619.

(13) (a) Wells, D. H., Jr.; Delgass, W. N.; Thomson, K. T. *J. Chem. Phys.* **2002**, *117*, 10597. (b) Weis, P. *Int. J. Mass Spectrom.* **2005**, *245*, 1.

(14) Taylor, K. J.; Pettiette-Hall, C. L.; Cheshnovski, O.; Smalley, R. E. *J. Chem. Phys.* **1992**, *96*, 3319.

(15) (a) Chesnovsky, O.; Pettiette, C. L.; Smalley, R. E. *Ion and Cluster Ion Spectroscopy and Structure*; Maier, J. P., Ed.; Elsevier: Amsterdam, 1988. (b) Jackschath, C.; Rabin, I.; Schulze, W. *Ber. Bunsen-Ges. Phys. Chem.* **1992**, *96*, 1200.

(16) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.

(17) Martin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408.

(18) Andrae, D.; Haeussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123.

(19) (a) Bishea, G. A.; Morse, M. D. *J. Chem. Phys.* **1991**, *95*, 5646. (b) Huber, K. P.; Herzberg, G. *Constants of Diatomic Molecules*; Van Nostrand Reinhold: New York, 1979. (c) Bauschlicher, C. W., Jr.; Langhoff, S. R.; Partridge, H. *J. Chem. Phys.* **1989**, *91*, 2412.

(20) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 533.

(21) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208.

(22) (a) Hilpert, K.; Gingerich, K. A. *Ber. Bunsen-Ges. Phys. Chem.* **1980**, *84*, 739. (b) Smard, B.; Hackett, P. A. *J. Mol. Spectrosc.* **1990**, *142*, 310.

(23) 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 C.02; Gaussian, Inc., Wallingford, CT, 2004.

(24) 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.; Bernhardsson, 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.; Köppl, 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, a package of ab initio programs.

(25) Kresse, G.; Furthmuller, J. *Phys. Rev. B* **1996**, *54*, 11169.

(26) Lee, S. J.; Chung, H. Y.; Kim, K. S. *Bull. Korean Chem. Soc.* **2004**, *25*, 1061.

(27) Diefenbach, M.; Kim, K. S. *J. Phys. Chem. B* **2006**, *110*, 21639.

(28) (a) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145. (b) Lee, H. M.; Suh, S. B.; Lee, J. Y.; Tarakeshwar, P.; Kim, K. S. *J. Chem. Phys.* **2000**, *112*, 9759.

(29) (a) Mendizabal, F.; Pyykkö, P. *Phys. Chem. Chem. Phys.* **2004**, *6*, 900. (b) Hess, B. A.; Kaldor, U. *J. Chem. Phys.* **2000**, *112*, 1809.

(30) Luo, W.; Pennycook, S. J.; Pantelides, S. T. *Nano Lett.* **2007**, *7*, 3134.

(31) Kim, W. Y.; Kim, K. S. *Nat. Nanotechnol.* **2008**, *3*, 408.

CT8003113

JCTC Journal of Chemical Theory and Computation

Quartic-Scaling Analytical Gradient of Quasidegenerate Scaled Opposite Spin Second-Order Perturbation Corrections to Single Excitation Configuration Interaction

Young Min Rhee,† David Casanova, and Martin Head-Gordon*

Department of Chemistry, University of California, Berkeley, California 94720, and Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720

Received November 25, 2008

Abstract: Quasidegenerate scaled second-order perturbation correction to single excitation configuration interaction (SOS-CIS(D_0)) is a viable method that can describe excited-state potential energy surfaces of various chemical systems both reliably and efficiently [*J. Chem. Phys.* **2008**, *128*, 164106]. In this work, its analytical gradient theory is developed and implemented into an efficient quartic-scaling algorithm. This low order scaling, as opposed to the traditional quintic scaling of various second-order perturbation methods, is attained by using the resolution-of-the-identity approximation and the Laplace transform. The efficiency of the method is demonstrated by calculating the excited-state gradients of molecules with varying sizes. The proposed gradient method will thus be useful in studying various chemical systems, ranging from finding the optimized stable geometry on the excited surface to elucidating interesting excited-state dynamics around the avoided crossing region.

I. Introduction

Obtaining an efficient and reliable theory for describing electronically excited states has been a long-standing research theme for many quantum chemists over many decades. For developing theories that can describe excitations from single-reference ground states, there have been a number of branches of endeavors: designing more reliable correction schemes for single excitation configuration interaction (CIS),[1] implementing linear response[2−4] or equation-of-motion theory[5,6] for various levels of coupled-cluster (CC)[7] expansions, and trying to eliminate errors in exchange-correlation description for the formally accurate time-dependent density functional theory (TDDFT).[8,9] Of these three branches, the correction scheme on CIS and the linear response scheme on truncated CC have much in common, primarily because they are both wave function-based theories.

In fact, developing efficient and reliable wave function-based excited-state theories is entering into a Renaissance in recent years. Of course, this is in part due to the advance of high performance computational platforms, where more costly calculations with wave function-based methods can be executed with a relative ease. However, recent endeavors in designing new algorithms and quantum chemical models and the subsequent successes have been the more important driving force for the changes. For example, the resolution-of-the-identity (RI) approximation,[10−15] which uses the density fitting[16−18] formulation by auxiliary basis expansion, has been widely adopted in recent developments, because it significantly reduces the computational cost associated with transforming various atomic orbital (AO) based electron repulsion integrals into molecular orbital (MO) based ones. Another important development has been the application of the Laplace transform algorithm proposed by Almlöf and co-workers.[19,20] Through this application, Scuseria and co-workers have shown that the traditional fifth-order scaling Møller−Plesset second-order perturbation scheme (MP2)[21] can be rendered into a linear scaling algorithm when applied to large enough molecular systems.[22] They have also shown that the MP2 theory can be efficiently applied to periodic systems through the use of the Laplace transform real-space formulation.[23] Furthermore, the RI and the Laplace

* Corresponding author e-mail: mhg@cchem.berkeley.edu.

† Present address: Department of Chemistry, Pohang University of Science and Technology, Pohang 790-784, Korea.

10.1021/ct800509z CCC: $40.75
Published on Web 04/14/2009

transform techniques have been recently combined to improve the efficiencies of the local MP2 model[24] and of the MP2 scheme for periodic systems.[25] It is also interesting to note the development of the spin component scaling (SCS) concept, which was pioneered by Grimme and co-workers[26,27] and rigorously interpreted by Szabados.[28] Relevant to the present work, Head-Gordon and co-workers have shown that the scaled opposite spin (SOS) version of perturbative corrections can improve the accuracy of both ground-state theory[29] (MP2) and excited-state theory[30] (perturbative doubles corrections for CIS; CIS(D)[31]). A similar behavior was also reported[32] for the quasidegenerate development of the perturbative correction scheme, CIS(D_0).[33] Hättig and co-workers have also reported[34] similar improvements in the robustness of truncated coupled-cluster theory CC2[35] through the use of these SCS and SOS approaches.

Among the above two scaling approaches, the SOS implementation has an important advantage: its computational demands for energy calculations only scale as fourth-order with respect to the molecular size when it is combined with the above-mentioned RI and Laplace transform techniques.[29,30,32] Achieving this scaling property, which is formally similar to the cost of mean field approaches such as CIS or TDDFT, is an important development because it allows us to apply the new methods to the calculation of much larger systems than what is practically feasible with conventional fifth-order scaling CIS(D)/CC2 or their SCS variants. In the same sense, attaining a similar efficiency in calculating the corresponding analytical gradient of the method will also be important, as the gradient is used in various situations of quantum chemical studies. In fact, for the ground-state SOS theory (SOS-MP2), it is already proven that such an efficient calculation is actually possible.[36]

In this work, we derive the analytical gradient of the fourth-order scaling quasidegenerate excited-state theory SOS-CIS(D_0). We prove that its gradient can also be obtained with a fourth-order scaling computational effort and propose an efficient algorithm for its actual computation. Then we apply the algorithm to various molecular systems and demonstrate that the gradient can actually be obtained very efficiently. The efficiency is shown to be strikingly good when it is compared with the case of the analytic gradient of the sister ground-state theory, SOS-MP2. As SOS-CIS(D_0) is an efficient and appropriate method for describing excited-state potential energy surfaces even near degeneracies, and as it does not suffer from the notorious problem for describing charge transfer excitations as in conventional TDDFT,[37−43] we expect that this gradient implementation will be useful in studying various chemical systems in their electronic excitations. Such an aspect is also discussed as a conclusion of this article.

II. Theory

In this section we will describe the formulation of the analytical gradient of SOS-CIS(D_0) theory, whose energy expression is given as[32,33]

$$\omega = \mathbf{b}\mathbf{A}^{(0)}_{SS}\mathbf{b} + c_T\mathbf{b}\mathbf{A}^{(2)}_{SS}\mathbf{b} + c_U\mathbf{b}\mathbf{A}^{(1)}_{SD}[\mathbf{D}^{(0)}_{DD}]^{-1}\mathbf{A}^{(1)}_{DS}\mathbf{b} \quad (1)$$

When we define the system Hamiltonian and its first-order fluctuation potential as $\hat{H}$ and $\hat{V}$, the response matrix elements can be expressed as

$$\mathbf{A}^{(0)}_{SS} = \langle\Phi_S|\hat{H}|\Phi_S\rangle \quad (2)$$

$$\mathbf{A}^{(2)}_{SS} = \langle\Phi_S|\hat{V}|\hat{T}_2^{OS}\Phi_S\rangle \quad (3)$$

$$\mathbf{A}^{(1)}_{SD} = \langle\Phi_S|\hat{V}|\Phi_D^{OS}\rangle \quad (4)$$

together with the well-known Møller−Plesset excitation operator $\hat{T}_2^{OS}$ in the opposite spin space. Here Φ_S represents a Slater determinant singly excited from the ground configuration Φ_0, while Φ_D^{OS} represents a determinant doubly excited in the opposite spin space. The expression for $\mathbf{A}^{(1)}_{DS}$ is obvious from eq 4, and the zeroth-order DD block is the diagonal eigenvalue difference matrix $\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j$. In addition, **b** represents the vector with single excitation amplitudes which is obtained as the eigenvector of the entire response matrix, $\mathbf{A}^{(0)}_{SS} + c_T\mathbf{A}^{(2)}_{SS} + c_U\mathbf{A}^{(1)}_{SD}[\mathbf{D}^{(0)}_{DD}]^{-1}\mathbf{A}^{(1)}_{DS}$. Finally, c_T and c_U represent the opposite spin component scaling parameters as defined previously.[30,32]

For future references, let us define the three terms on the right-hand side of eq 1 as ω_0, ω_T, and ω_U, respectively

$$\omega = \omega_0 + \omega_T + \omega_U \quad (5)$$

Of the three, ω_0 requires only fourth-order scaling contractions between $\mathbf{A}^{(0)}_{SS}$ and **b**, and we will use exact integrals without any approximations for this term. This is similar to the case of evaluating electron repulsion integrals (ERIs) in Hartree−Fock (HF) self-consistent field (SCF) solution, where exact integrals are used without the RI approximation. For ω_T and ω_U, ERIs are evaluated with the RI approximation[10−16,44]

$$(pq|rs)_{RI} = \sum_{PQ}(pq|P)(P|Q)^{-1}(Q|rs) = \sum_R B^R_{pq}B^R_{rs} \quad (6)$$

with the **B** matrix defined as

$$B^R_{pq} = \sum_P (pq|P)(P|R)^{-1/2} \quad (7)$$

together with an auxiliary basis set $\{P\}$. With the above explanation, one can easily infer where RI approximations are adopted, and we will not explicitly use the subscript "RI" in the following equations for the sake of simplicity.

Because **b** is obtained as the eigenvector of the symmetrized response matrix,[32] the first derivative of ω with respect to some parameter x can be obtained as

$$\omega^{(x)} = \mathbf{b}[\mathbf{A}^{(0)}_{SS}]^{(x)}\mathbf{b} + c_T\mathbf{b}[\mathbf{A}^{(2)}_{SS}]^{(x)}\mathbf{b} + c_U\mathbf{b}[\mathbf{A}^{(1)}_{SD}(\mathbf{D}^{(0)}_{DD})^{-1}\mathbf{A}^{(1)}_{DS}]^{(x)}\mathbf{b} \quad (8)$$

The gradient expression originating from ω_0 term is identical to the CIS case, and readers are referred to the detailed derivations in refs 45 and 46.

As in the case of the sister ground-state theory SOS-MP2, the quartic-scaling SOS-CIS(D_0) utilizes the Laplace transform of the energy denominator[19,29,30,32]

$$\frac{1}{\Delta^{ab}_{ij}} = \sum_t \rho_t \exp(-\Delta^{ab}_{ij}t) \quad (9)$$

with ρ_t denoting the weight at any given quadrature point t, and $\mathbf{\Delta}$ representing the diagonal energy difference tensor $\Delta^{ab}_{ij} = \varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j$. Therefore, to obtain the expression of the

analytical gradient of both SOS-MP2 and SOS-CIS(D_0), one needs to derive the first derivative of $e^{-\mathbf{\Delta}t}$. Because a tensor $\mathbf{\Delta}$ and its derivative $\mathbf{\Delta}^{(x)}$ do not commute unless both are diagonal, the following simple formulation is not generally satisfied

$$[e^{-\mathbf{\Delta}t}]^{(x)} \neq -te^{-\mathbf{\Delta}t}\mathbf{\Delta}^{(x)} \tag{10}$$

One approach to satisfy the above equality is to restrict the orbitals to be canonical by setting the Fock matrix ε_{pq} diagonal even after the perturbation "x"[47]

$$\varepsilon_{pq}^{(x)} = \varepsilon_{pq}^{(x)}\delta_{pq} \tag{11}$$

Indeed, Lochan et al. developed the quartic-scaling SOS-MP2 gradient theory through the use of this canonical representation.[36]

Even though $\mathbf{\Delta}^{(x)}$ may not be diagonal in general, we should note from Brillouin's theorem that it is still block-diagonal.[48] In fact, as long as $\mathbf{\Delta}$ is diagonal and $\mathbf{\Delta}^{(x)}$ is block-diagonal, one can show that the following equation is satisfied (see the Appendix for its detailed proof)

$$\begin{aligned}[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'} = &-t\varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'}e^{-\Delta^{ab}_{ij}t}h(\Delta^{a'}_{a}t) - \\ &t\varepsilon^{(x)}_{bb'}\delta_{ii'}\delta_{jj'}\delta_{aa'}e^{-\Delta^{ab}_{ij}t}h(\Delta^{b'}_{b}t) + t\varepsilon^{(x)}_{ii'}\delta_{ii'}\delta_{aa}\delta_{bb}e^{-\Delta^{ab}_{ij}t}h(\Delta^{i}_{i}t) + \\ &t\varepsilon^{(x)}_{jj'}\delta_{ii'}\delta_{aa}\delta_{bb}e^{-\Delta^{ab}_{ij}t}h(\Delta^{j}_{j}t)\end{aligned} \tag{12}$$

with $h(x) = (1 - e^{-x})/x$ and $\Delta^q_p = \varepsilon_q - \varepsilon_p$. With this more generalized equation for $\mathbf{\Delta}^{(x)}$, one can use the symmetric representation of the occupied-occupied and virtual-virtual orbital responses[49,50]

$$U^x_{ij} = -\frac{1}{2}S^x_{ij} \tag{13a}$$

$$U^x_{ab} = -\frac{1}{2}S^x_{ab} \tag{13b}$$

where the $\mathbf{S}^x$ matrices correspond to the derivatives of the overlap integrals. Curious readers can rederive SOS-MP2 gradient using this symmetric representation and can show that the final expressions are identical to the ones obtained with the canonical representation. Besides this interesting complementary nature, the benefit of using the symmetric representation presented here is the associated simplicity in the derivation of the analytical gradient. Accordingly, we will exclusively use the symmetric representation in conjunction with eq 12.

A. Gradient from ω_T Term. In this section we derive the expression for the analytical gradient of the second term in eq 1. Following the conventional orbital designations[30] (i, j, ... for occupied; a, b, ... for virtual; overbar for β-spin), the detailed expression for this term can be given as

$$\begin{aligned}\omega_T = &-a^{\bar{b}\bar{c}}_{\bar{j}\bar{k}}b^{\bar{a}}_{\bar{i}}b^{\bar{b}}_{\bar{i}}(\bar{j}\bar{a}|kc) - a^{\bar{b}\bar{c}}_{\bar{j}\bar{k}}b^{\bar{a}}_{i}b^{\bar{a}}_{\bar{j}}(\bar{i}\bar{b}|kc) + \\ &a^{\bar{b}\bar{c}}_{\bar{j}\bar{k}}b^{a}_{i}b^{\bar{b}}_{\bar{j}}(ik||ac) + a^{\bar{b}\bar{c}}_{\bar{j}\bar{k}}b^{\bar{a}}_{i}b^{\bar{b}}_{\bar{j}}(\bar{i}\bar{a}|kc) - \\ &a^{b\bar{c}}_{j\bar{k}}b^{a}_{i}b^{b}_{j}(ja|\bar{k}\bar{c}) - a^{b\bar{c}}_{j\bar{k}}b^{a}_{i}b^{a}_{j}(ib|\bar{k}\bar{c}) + \\ &a^{b\bar{c}}_{j\bar{k}}b^{\bar{a}}_{i}b^{b}_{\bar{j}}(\bar{i}\bar{k}||\bar{a}\bar{c}) + a^{b\bar{c}}_{j\bar{k}}b^{a}_{i}b^{b}_{j}(ia|\bar{k}\bar{c})\end{aligned} \tag{14}$$

Because the closed-shell expression can be easily obtained from this general spin-unrestricted description by setting $b^{\bar{a}}_{\bar{i}} = b^{a}_{i}$ ($b^{\bar{a}}_{\bar{i}} = -b^{a}_{i}$) for singlet (triplet) excitations, we will only present the spin-unrestricted equations. By applying eq 12, the gradient of its first term can be formulated as

$$\begin{aligned}\left[-a^{\bar{b}\bar{c}}_{\bar{j}\bar{k}}b^{\bar{a}}_{\bar{i}}b^{\bar{b}}_{\bar{i}}(\bar{j}\bar{a}|kc)\right]^{(x)} \doteq &\sum_t \rho_t\Big[(\bar{j}\bar{b}|kc)^{(x)}\widetilde{W}^{Q}_{\bar{b}\bar{j}}\widetilde{B}^{Q}_{ck} + \widetilde{B}^{P}_{\bar{b}\bar{j}}\widetilde{B}^{P}_{ck}P_{\bar{a}\bar{b}}(\bar{j}\bar{a}|kc)^{(x)}\Big] + \\ &\sum_t \rho_t P_{\bar{a}\bar{b}}(\bar{j}\bar{a}|kc)[-\varepsilon^{(x)}_{\bar{b}\bar{d}}\exp(-\Delta^{\bar{d}c}_{\bar{j}\bar{k}}t)\widetilde{h}^{\bar{b}}_{\bar{d}}t(\bar{j}\bar{d}|kc) - \\ &\varepsilon^{(x)}_{cd}\exp(-\Delta^{\bar{b}d}_{\bar{j}k}t)\widetilde{h}^{c}_{d}t(\bar{j}\bar{b}|kd) + \varepsilon^{(x)}_{\bar{j}\bar{l}}\exp(-\Delta^{\bar{b}c}_{\bar{l}k}t)\widetilde{h}^{\bar{l}}_{\bar{j}}t(\bar{l}\bar{b}|kc) + \\ &\varepsilon^{(x)}_{kl}\exp(-\Delta^{\bar{b}c}_{\bar{j}l}t)\widetilde{h}^{l}_{k}t(\bar{j}\bar{b}|lc)]\end{aligned} \tag{15}$$

Because $\mathbf{b}^{(x)}$ does not appear in the overall derivative of the SOS-CIS(D_0) energy according to eq 8, we can safely omit any terms involving $\mathbf{b}^{(x)}$. The dot over the equality sign in the above equation explicitly remarks this omission. In reaching eq 15, we have used the definition

$$\tilde{h}^q_p = h(\Delta^q_p t) = \frac{1 - e^{-\Delta^q_p t}}{\Delta^q_p t} \tag{16}$$

together with its index exchange property given as

$$\tilde{h}^p_q = \tilde{h}^q_p e^{-\Delta^p_q t} \tag{17}$$

The definitions of other intermediates adopted in eq 15 and any subsequent equations can be found in Table 1.

The derivatives of the remaining terms from eq 14 can be obtained in a similar fashion. After some algebra, the analytical gradient of ω_T can be expressed in a rather compact form

$$\begin{aligned}[\omega_T]^{(x)} \doteq &2\Pi^P_{ai}(ai|P)^x + 2\Pi^P_{\bar{a}\bar{i}}(\bar{a}\bar{i}|P)^x - \gamma^{\Pi}_{RS}(R|S)^x + \\ &W^{\Pi}_{ab}S^x_{ab} + W^{\Pi}_{\bar{a}\bar{b}}S^x_{\bar{a}\bar{b}} + 2W^{\Pi}_{ai}S^x_{ai} + 2W^{\Pi}_{\bar{a}\bar{i}}S^x_{\bar{a}\bar{i}} + W^{\Pi}_{ij}S^x_{ij} + \\ &W^{\Pi}_{\bar{i}\bar{j}}S^x_{\bar{i}\bar{j}} + 2L^{\Pi}_{ai}U^x_{ai} + 2L^{\Pi}_{\bar{a}\bar{i}}U^x_{\bar{a}\bar{i}} + P^{\Pi}_{ab}\varepsilon^{(x)}_{ab} + P^{\Pi}_{\bar{a}\bar{b}}\varepsilon^{(x)}_{\bar{a}\bar{b}} + \\ &P^{\Pi}_{ij}\varepsilon^{(x)}_{ij} + P^{\Pi}_{\bar{i}\bar{j}}\varepsilon^{(x)}_{\bar{i}\bar{j}}\end{aligned} \tag{18}$$

with $(ai|P)^x$ denoting the atomic orbital derivative of the 3-centered 2-electron repulsion integral within the conventional RI approximation with an auxiliary basis set $\{P\}$. The 3-centered 2-particle density matrix (3C-2PDM) is defined as

$$\begin{aligned}\Pi^P_{ai} &= \Lambda^P_{ai} + \frac{1}{2}\Omega^P_{ai} + \frac{1}{2}\mathrm{X}^P_{ai} \\ &= \sum_t \rho_t\left(\tilde{\Lambda}^P_{ai} + \frac{1}{2}\tilde{\Omega}^P_{ai}\right) + \frac{1}{2}\mathrm{X}^P_{ai}\end{aligned} \tag{19}$$

With the intermediate 2PDMs given as

$$\tilde{\Lambda}^P_{ai} = \frac{1}{2}(P|R)^{-1/2}({}^{\beta}\tilde{U}_{RQ} - {}^{\beta}\tilde{S}_{RQ} + {}^{\beta}\tilde{U}_{QR} - {}^{\beta}\tilde{S}_{QR})\tilde{B}^Q_{ai} \tag{20}$$

$$\tilde{\Omega}^P_{ai} = {}^{\beta}\tilde{x}_{PQ}(\tilde{W}^Q_{ai} - \tilde{Q}^Q_{ai}) - (P|Q)^{-1/2\,\beta}\tilde{h}^Q\tilde{b}^a_i - (P|Q)^{-1/2}\tilde{I}_{ai}{}^{\beta}\tilde{f}^Q \tag{21}$$

$$\begin{aligned}\mathrm{X}^P_{ai} = &-P_{ab}\Gamma^P_{bi} + \Gamma^P_{aj}P_{ij} - b^a_i({}^{\alpha}g^P + {}^{\beta}g^P) - \\ &(P|R)^{-1/2}({}^{\alpha}f^R + {}^{\beta}f^R)H_{ai} + b^a_kC^P_{ck}H_{ci} + b^c_iC^P_{ck}H_{ak}\end{aligned} \tag{22}$$

Here, $\mathbf{\Gamma}$ is the SOS-MP2 ground-state 2PDM, which can be obtained with[36]

$$\Gamma^P_{ai} = -\sum_t \rho_t {}^{\beta}\tilde{x}_{PQ}\tilde{B}^Q_{ai} \tag{23}$$

The 1-particle density matrix (1PDM) corrections can be derived as

$$P^{\Pi}_{ab} = -\sum_t \rho_t \tilde{h}^a_b t(\tilde{\Lambda}^Q_{bj} + \tilde{\Omega}^Q_{bj})(aj|Q) \tag{24}$$

$$P^{\Pi}_{ij} = \sum_t \rho_t \tilde{h}^j_i t(\tilde{\Lambda}^Q_{bj} + \tilde{\Omega}^Q_{bj})(ib|Q) \tag{25}$$

In arriving at these equations, we have used

$$\tilde{h}^b_a t({}^{\beta}\tilde{U}_{PQ} - {}^{\beta}\tilde{S}_{PQ})\tilde{B}^P_{aj}B^Q_{bj} = \tilde{h}^a_b t({}^{\beta}\tilde{U}_{QP} - {}^{\beta}\tilde{S}_{QP})\tilde{B}^P_{bj}B^Q_{aj} \quad \text{(no sum in } a,b) \tag{26a}$$

$$\tilde{h}^j_i t({}^{\beta}\tilde{U}_{PQ} - {}^{\beta}\tilde{S}_{PQ})\tilde{B}^P_{aj}B^Q_{ai} = \tilde{h}^j_i t({}^{\beta}\tilde{U}_{QP} - {}^{\beta}\tilde{S}_{QP})\tilde{B}^P_{ai}B^Q_{aj} \quad \text{(no sum in } i,j) \tag{26b}$$

together with the index symmetries of $\varepsilon^{(x)}_{ab}$ and $\varepsilon^{(x)}_{ij}$.

The energy weighted density matrices (**W**) and the Lagrangian (**L**) can be obtained in a manner similar to the SOS-MP2 or RI-MP2 gradients by defining two intermediate matrices

$$L^{\Pi,1}_{a\mu} = \Pi^Q_{aj}(\mu j|Q) \tag{27}$$

$$L^{\Pi,2}_{\nu i} = \Pi^Q_{aj}(a\nu|Q) \tag{28}$$

with the usual designation of μ, ν,... for the atomic orbitals. With these definitions, the working expressions of **W** and **L** follow as

$$W^{\Pi}_{ap} = -L^{\Pi,1}_{a\mu}C_{\mu p} \tag{29}$$

$$W^{\Pi}_{ij} = -L^{\Pi,2}_{\nu i}C_{\nu j} \tag{30}$$

$$L^{\Pi}_{ai} = -L^{\Pi,1}_{a\mu}C_{\mu i} + L^{\Pi,2}_{\nu i}C_{\nu a} \tag{31}$$

The quantities in the β-space are defined analogously and will not be elaborated here. Finally, the 2-centered 2-particle density matrix (2C-2PDM) is obtained as

$$\gamma^{\Pi}_{RS} = \sum_t \rho_t[2{}^{\beta}\tilde{e}_{RQ}{}^{\alpha}\tilde{x}_{SQ} + 2{}^{\alpha}\tilde{e}_{RQ}{}^{\beta}\tilde{x}_{SQ} - (S|T)^{-1/2\alpha}\tilde{h}^{T\beta}\tilde{f}^Q(Q|R)^{-1/2} - (S|T)^{-1/2\beta}\tilde{h}^{T\alpha}\tilde{f}^Q(Q|R)^{-1/2} - ({}^{\alpha}g^S + {}^{\beta}g^S)({}^{\alpha}f^Q + {}^{\beta}f^Q)(Q|R)^{-1/2} + H_{bk}C^R_{bj}C^S_{ck}b^c_j + H_{\bar{b}\bar{k}}C^R_{\bar{b}j}C^S_{\bar{c}\bar{k}}b^{\bar{c}}_{\bar{j}}] \tag{32}$$

B. Gradient from ω_U Term. Making use of the Laplace transform (eq 9), ω_U can be expressed as

$$\omega_U = -\sum_t \rho_t u^{\bar{a}\bar{b}}_{\bar{i}\bar{j}}\exp(-\Delta^{\bar{a}\bar{b}}_{\bar{i}\bar{j}}t)u^{\bar{a}\bar{b}}_{\bar{i}\bar{j}} \tag{33}$$

$$u^{\bar{a}\bar{b}}_{\bar{i}\bar{j}} = (\bar{a}\bar{c}|bj)b^{\bar{c}}_{\bar{i}} + (\bar{a}\bar{i}|bc)b^c_j - (\bar{k}\bar{i}|bj)b^{\bar{a}}_{\bar{k}} - (kj|\bar{a}\bar{i})b^b_k \tag{34}$$

and with the help of eq 12, its derivative can be easily formulated as

$$[\omega_U]^{(x)} \doteq -\sum_t 2\rho_t u^{\bar{a}\bar{b}}_{\bar{i}\bar{j}}\exp(-\Delta^{\bar{a}\bar{b}}_{\bar{i}\bar{j}}t)[(\bar{a}\bar{c}|bj)^{(x)}b^{\bar{c}}_{\bar{i}} + (\bar{a}\bar{i}|bc)^{(x)}b^c_j - (\bar{k}\bar{i}|bj)^{(x)}b^{\bar{a}}_{\bar{k}} - (kj|\bar{a}\bar{i})^{(x)}b^b_k] - \sum_t \rho_t u^{\bar{a}\bar{b}}_{\bar{i}\bar{j}}\exp(-\Delta^{\bar{a}\bar{b}}_{\bar{i}\bar{j}}t)[-\varepsilon^{(x)}_{\bar{a}\bar{d}}\tilde{h}^{\bar{d}}_{\bar{a}}tu^{\bar{d}\bar{b}}_{\bar{i}\bar{j}} - \varepsilon^{(x)}_{bd}\tilde{h}^d_b tu^{\bar{a}\bar{d}}_{\bar{i}\bar{j}} + \varepsilon^{(x)}_{\bar{i}\bar{l}}\tilde{h}^{\bar{i}}_{\bar{l}}tu^{\bar{a}\bar{b}}_{\bar{l}\bar{j}} + \varepsilon^{(x)}_{jl}\tilde{h}^j_l tu^{\bar{a}\bar{b}}_{\bar{i}\bar{l}}] \tag{35}$$

Similar to $[\omega_T]^{(x)}$, this can be straightforwardly transformed into a compact form after some algebra

$$[\omega_U]^{(x)} \doteq 2\mathrm{M}^P_{aj}b^b_j(ab|P)^x + 2\mathrm{M}^P_{\bar{a}\bar{j}}b^{\bar{b}}_{\bar{j}}(\bar{a}\bar{b}|P)^x - 2\mathrm{M}^P_{bj}b^b_i(ij|P)^x - 2\mathrm{M}^P_{\bar{b}\bar{j}}b^{\bar{b}}_{\bar{i}}(\bar{i}\bar{j}|P)^x + 2\mathrm{N}^P_{ai}(ai|P)^x + 2\mathrm{N}^P_{\bar{a}\bar{i}}(\bar{a}\bar{i}|P)^x - \gamma^{\mathrm{D}}_{RS}(R|S)^x + W^{\mathrm{D}}_{ab}S^x_{ab} + W^{\mathrm{D}}_{\bar{a}\bar{b}}S^x_{\bar{a}\bar{b}} + 2W^{\mathrm{D}}_{ai}S^x_{ai} + 2W^{\mathrm{D}}_{\bar{a}\bar{i}}S^x_{\bar{a}\bar{i}} + W^{\mathrm{D}}_{ij}S^x_{ij} + W^{\mathrm{D}}_{\bar{i}\bar{j}}S^x_{\bar{i}\bar{j}} + 2L^{\mathrm{D}}_{ai}U^x_{ai} + 2L^{\mathrm{D}}_{\bar{a}\bar{i}}U^x_{\bar{a}\bar{i}} + P^{\mathrm{D}}_{ab}\varepsilon^{(x)}_{ab} + P^{\mathrm{D}}_{\bar{a}\bar{b}}\varepsilon^{(x)}_{\bar{a}\bar{b}} + P^{\mathrm{D}}_{ij}\varepsilon^{(x)}_{ij} + P^{\mathrm{D}}_{\bar{i}\bar{j}}\varepsilon^{(x)}_{\bar{i}\bar{j}} \tag{36}$$

It is easy to show that 3C-2PDMs on each quadrature point are obtained as

$$\tilde{\mathrm{M}}^P_{ai} = -{}^{\beta}\tilde{x}_{PQ}\tilde{D}^Q_{ai} - {}^{\beta}\tilde{y}_{PQ}\tilde{B}^Q_{ai} \tag{37}$$

$$\tilde{\mathrm{N}}^P_{ai} = -{}^{\beta}\tilde{w}_{PQ}\tilde{D}^Q_{ai} - {}^{\beta}\tilde{z}_{PQ}\tilde{B}^Q_{ai} \tag{38}$$

with their obvious counterparts in the β-space. The working expression for the 2C-2PDM is

$$\gamma^{\mathrm{D}}_{RS} = -2\sum_t \rho_t[{}^{\beta}\tilde{z}_{RQ}{}^{\alpha}\tilde{x}_{SQ} + {}^{\beta}\tilde{y}_{RQ}{}^{\alpha}\tilde{w}_{SQ} + {}^{\beta}\tilde{x}_{RQ}{}^{\alpha}\tilde{z}_{SQ} + {}^{\beta}\tilde{w}_{RQ}{}^{\alpha}\tilde{y}_{SQ}] \tag{39}$$

and the 1PDM corrections are

$$P^{\mathrm{D}}_{ab} = -\sum_t \rho_t \tilde{h}^a_b t(\tilde{\mathrm{M}}^Q_{bj}A^Q_{aj} + \tilde{\mathrm{N}}^Q_{bj}(aj|Q)) \tag{40}$$

$$P^{\mathrm{D}}_{ij} = \sum_t \rho_t \tilde{h}^j_i t(\tilde{\mathrm{M}}^Q_{bj}A^Q_{bi} + \tilde{\mathrm{N}}^Q_{bj}(ib|Q)) \tag{41}$$

also with their obvious counterparts in the β-space. If we define the following two matrices

$$L^{\mathrm{D},1}_{a\mu} = \mathrm{N}^Q_{aj}(\mu j|Q) + \mathrm{M}^Q_{aj}b^c_j(c\mu|Q) + \mathrm{M}^Q_{bj}b^a_j(b\mu|Q) \tag{42}$$

$$L^{\mathrm{D},2}_{\nu i} = \mathrm{N}^Q_{bi}(b\nu|Q) - \mathrm{M}^Q_{bj}b^b_i(\nu j|Q) - \mathrm{M}^Q_{bi}b^b_j(\nu j|Q) \tag{43}$$

the energy-weight densities and the Lagrangian, $\mathbf{L}^{\mathrm{D}}$ and $\mathbf{W}^{\mathrm{D}}$, can be obtained analogously to eqs 29−31.

C. Efficient Usage of 3C-2PDMs. For an efficient evaluation of the intermediate **L** and 3-centered force (3CF) terms with $(pq|R)^x$, it is customary to use both the primitive and the back-transformed 3C-2PDMs

$$L^1_{a\mu} = \Gamma^Q_{aj}[(\mu\nu|Q)C_{\nu j}] \tag{44}$$

$$L^2_{\nu i} = \Gamma^Q_{\mu i}(\mu\nu|Q) = [\Gamma^Q_{ai}C_{\mu a}](\mu\nu|Q) \tag{45}$$

$$\Gamma^Q_{ai}(ai|Q)^x = \Gamma^Q_{\mu i}(\mu\nu|Q)^x C_{\nu i} = [\Gamma^Q_{ai}C_{\mu a}](\mu\nu|Q)^x C_{\nu i} \tag{46}$$

Here, the square brackets explicitly show the necessary

Table 1. Definitions of Various Intermediates for Evaluating SOS-CIS(D_0) Gradient[a-c]

Laplace quadrature independent	Laplace quadrature dependent
$P_{ij} = -b_i^a b_j^a$	$\tilde{B}_{ai}^P = B_{ai}^P e^{-\varepsilon_a t} e^{\varepsilon_i t}$
$P_{ab} = b_i^a b_i^b$	$\tilde{D}_{ai}^P = D_{ai}^P e^{-\varepsilon_a t} e^{\varepsilon_i t}$
$B_{aj}^P = (aj\|R)(R\|R)^{-1/2}$	$\tilde{W}_{ai}^P = W_{ai}^P e^{-\varepsilon_a t} e^{\varepsilon_i t}$
$C_{aj}^P = (aj\|R)(R\|P)^{-1}$	$\tilde{Q}_{ai}^P = Q_{ai}^P e^{-\varepsilon_a t} e^{\varepsilon_i t}$
$D_{ai}^P = B_{ab}^P b_i^b - B_{ji}^P b_j^a$	$\tilde{f}^P = b_i^a \tilde{B}_{ai}^P,\ \bar{f}^P = \tilde{f}^R (P\|R)^{-1/2}$
$A_{ai}^P = D_{ai}^R (R\|P)^{-1/2}$	$\tilde{I}_{ai} = I_{ai} e^{-\varepsilon_a t} e^{\varepsilon_i t}$
$V_{i(j)}^P = b_j^a B_{ai}^R$	$\tilde{X}_{PQ} = B_{ai}^P \tilde{B}_{ai}^Q,\ \tilde{x}_{PQ} = (P\|R)^{-1/2} \tilde{X}_{RQ}$
$W_{aj}^P = P_{ab} B_{bj}^P$	$\tilde{Y}_{PQ} = \tilde{D}_{ai}^P B_{ai}^Q = D_{ai}^P \tilde{B}_{ai}^Q$
$Q_{ai}^P = B_{aj}^P P_{ji}$	$\tilde{Z}_{PQ} = D_{ai}^P \tilde{D}_{ai}^Q,\ \tilde{z}_{PQ} = (P\|R)^{-1/2} \tilde{Z}_{QR}$
$f^P = b_i^a B_{ai}^P$	$\tilde{U}_{PQ} = W_{ai}^P \tilde{B}_{ai}^Q = \tilde{W}_{ai}^P B_{ai}^Q$
$J_{ai} = ({}^{\alpha}f^P + {}^{\beta}f^P) B_{ai}^P$	$\tilde{S}_{PQ} = Q_{ai}^P \tilde{B}_{ai}^Q = \tilde{Q}_{ai}^P B_{ai}^Q$
$K_{ai} = B_{aj}^P b_j^b B_{bi}^P$	$\tilde{y}_{PQ} = (P\|R)^{-1/2} \tilde{Y}_{QR}$
$I_{ai} = J_{ai} - K_{ai}$	$\tilde{w}_{PQ} = (P\|R)^{-1/2} \tilde{Y}_{RQ}$
$H_{ai} = \sum_t \rho_t^{\beta} \tilde{f}^P \tilde{B}_{ai}^P$	$\tilde{e}_{PQ} = {}^1/_2 (P\|R)^{-1/2} (\tilde{U}_{RQ} + \tilde{U}_{QR} - \tilde{S}_{RQ} - \tilde{S}_{QR})$
$g^P = \sum_t \rho_t^{\beta} \tilde{x}_{PQ}^a \tilde{f}^Q = H_{ai} C_{ai}^P$	$\tilde{h}^P = \tilde{I}_{bj} B_{bj}^P = I_{bj} \tilde{B}_{bj}^P,\ \bar{h}^P = \tilde{h}^R (P\|R)^{-1/2}$

[a] Only the terms in the α-space are presented. [b] Unless ambiguous, spin designations are omitted for simplicity. [c] Repeated indices implicitly mean summations except with energy weightings, $e^{\pm \varepsilon_p t}$.

contractions. This way, any computation or storage of matrices with the size $\sim VNX$ can be avoided, leaving only much smaller quantities with size $\sim ONX$. (Note that O represents the number of occupied orbitals; V, no. of virtual orbitals; N, no, of basis functions; X, no. of auxiliary basis functions.)

In SOS-CIS(D_0) gradient, the terms involving $\mathbf{\Pi}$ and $\mathbf{N}$ can be trivially evaluated in the same manner. For efficiency, the sum of these two 2PDMs can be used to invoke these $\mathbf{L}$ and 3CF evaluations. The second and third terms of eq 43 can actually be evaluated together with $\mathbf{\Pi}$ and $\mathbf{N}$ by defining

$$\mathrm{M}'^{Q}_{\mu i} = \mathrm{M}^{Q}_{bj} b_i^b C_{\mu j} + \mathrm{M}^{Q}_{bi} b_j^b C_{\mu j} \tag{47}$$

and

$$L^{\mathrm{I},1}_{a\mu} = (c_T \Pi^{Q}_{aj} + c_U \mathrm{N}^{Q}_{aj})[(\mu\nu|Q) C_{\nu j}] \tag{48}$$

$$L_{vi}^{\mathrm{I},2} = (c_T\Pi_{\mu i}^{Q} + c_U\mathrm{N}_{\mu i}^{Q} - c_U\mathrm{M}'^{Q}_{\mu i})(\mu\nu|Q) \quad (49)$$

Here, c_T and $c_U$refer to the spin component scaling parameters presented in eq 1. Using $\mu\nu$ index exchange symmetry in $(\mu\nu|Q)^x$, it can be shown that the 3CF terms involving $(ai|Q)^x$ and $(ij|Q)^x$ can also be obtained with

$$(3\mathrm{CF})_{\mathrm{I}}^{x} = (2c_T\Pi_{\mu i}^{Q} + 2c_U\mathrm{N}_{\mu i}^{Q} - c_U\mathrm{M}'^{Q}_{\mu i})(\mu\nu|Q)^x C_{\nu i} \quad (50)$$

It is interesting to note that the same approach can be employed to efficiently evaluate the remaining **L** and 3CF terms. For this purpose, we need to define a dressed coefficient

$$D_{vj} = C_{vb}b_j^b \quad (51)$$

and generate the following intermediate **L** matrices

$$L_{a\mu}^{\mathrm{II},1} = c_U\mathrm{M}_{aj}^{Q}[(\mu\nu|Q)D_{vj}] \quad (52)$$

$$L_{\nu i}^{\mathrm{II},2} = c_U\mathrm{M}_{\mu i}^{Q}(\mu\nu|Q) = c_U[\mathrm{M}_{ai}^{Q}C_{\mu a}](\mu\nu|Q) \quad (53)$$

Likewise, it is easy to show that the 3CF terms involving $(ab|Q)^x$ is given as

$$(3\mathrm{CF})_{\mathrm{II}}^{x} = 2c_U\mathrm{M}_{\mu i}^{Q}(\mu\nu|Q)^x D_{\nu i} \quad (54)$$

Then, the final **L** matrices from both $[\omega_T]^{(x)}$ and $[\omega_U]^{(x)}$ can be constructed as

$$L_{a\mu}^{1} = L_{a\mu}^{\mathrm{I},1} + L_{a\mu}^{\mathrm{II},1} + L_{\mu i}^{\mathrm{II},2}b_i^a \quad (55)$$

$$L_{\nu i}^{2} = L_{\nu i}^{\mathrm{I},2} \quad (56)$$

Of course, the final 3CF will be the sum of the two contributions, $(3\mathrm{CF})_{\mathrm{I}}$ and $(3\mathrm{CF})_{\mathrm{II}}$.

D. Finalizing the Gradient. Up to this point, we have seen how 2PDMs and Lagrangians are built together with the unrelaxed 1PDMs. After obtaining these, the gradient can be calculated in the same manner as in many other quantum chemical methods.[14,44,46,51,52] First, to compute the relaxed 1PDM, the **Z**-vector equation[53] is solved such that

$$\mathbf{AZ} = \mathbf{L} \quad (57)$$

is satisfied together with the usual orbital Hessian **A**.[54,55] Even though the ground- and the excited-state contributions to 1PDM can be calculated separately along with two separate **Z**-vectors as their individual relaxations in theory, it is practically more desirable to compute the overall 1PDM and its relaxation by using the combined Lagrangian.

After this, the final construction of the gradient can be performed using the well-known expression

$$E_{\mathrm{SOS\text{-}CIS(D_0)}}^{(x)} = P_{\mu\nu}^{\mathrm{tot}}H_{\mu\nu}^{x} + W_{\mu\nu}^{\mathrm{tot}}S_{\mu\nu}^{x} + 2\Gamma_{\mu\nu}^{\mathrm{tot};P}(\mu\nu|P)^x - \gamma_{RS}^{\mathrm{tot}}(R|S)^x + \Gamma_{\mu\nu\lambda\sigma}(\mu\nu|\lambda\sigma)^x \quad (58)$$

Here, we have used "tot" to denote that the density matrices should include contributions from all relevant energy terms (HF, SOS-MP2, and ω) with appropriate scaling parameters. As the explicit expressions of the terms from ω have been described in this section, and as the terms from other contributions are already well-known,[36,46,56] reiterating the formulation of these terms will be only laborious. Thus, we will leave the details up to the readers. Finally, $\Gamma_{\mu\nu\lambda\sigma}$ represents the separable 2PDM given as

$$\Gamma_{\mu\nu\lambda\sigma} = \left(\frac{1}{2}P_{\mu\nu}^{\mathrm{HF}} + P_{\mu\nu}^{\mathrm{corr}}\right)P_{\lambda\sigma}^{\mathrm{HF}} - \left(\frac{1}{2}P_{\mu\sigma}^{\mathrm{HF}} + P_{\mu\sigma}^{\mathrm{corr}}\right)P_{\lambda\nu}^{\mathrm{HF}} + T_{\mu\nu}T_{\lambda\sigma} - T_{\mu\sigma}T_{\lambda\nu} \quad (59)$$

with HF 1PDM $P_{\mu\nu}^{\mathrm{HF}}$, post-HF density correction $P_{\mu\nu}^{\mathrm{corr}}$, and CIS dressed density-like matrix $T_{\mu\nu}$ defined as[46]

$$T_{\mu\nu} = C_{\mu i}b_i^a C_{\nu a} \quad (60)$$

E. Avoiding Potential Divergence of Exponentials. The factor $\tilde{h}_p^q$ given in eq 16 contains exponential functions with 1-electron orbital energies as their arguments. This may lead

Chart 1. Algorithm of Laplace Loop I and Its CPU and IO Costs

		Disk IO cost	CPU cost
	Loop over t ($t \in$ quadrature points, size: T)		
	Loop over i ($i \in$ OCC)		
1	Read B_{ai}^{P}, D_{ai}^{P}, $E_{ai}^{P} = W_{ai}^{P} - Q_{ai}^{P}$	$3OVXT$	
2	$\tilde{B}_{ai}^{P} = B_{ai}^{P}e^{-(\varepsilon_a-\varepsilon_i)t}$, $\tilde{D}_{ai}^{P} = D_{ai}^{P}e^{-(\varepsilon_a-\varepsilon_i)t}$		
3	$\tilde{f}^{P}$ += $\sum_a b_i^a \tilde{B}_{ai}^{P}$		
4	$\tilde{h}^{P}$ += $\sum_a I_{ai}\tilde{B}_{ai}^{P}$		
5	$\tilde{X}_{PQ}$ += $\sum_a B_{ai}^{P}\tilde{B}_{ai}^{Q}$		OVX^2T
6	$\tilde{Y}_{PQ}$ += $\sum_a D_{ai}^{P}\tilde{B}_{ai}^{Q}$		OVX^2T
7	$\tilde{Z}_{PQ}$ += $\sum_a D_{ai}^{P}\tilde{D}_{ai}^{Q}$		OVX^2T
8	$(\tilde{U}_{PQ} - \tilde{S}_{PQ})$ += $\sum_a E_{ai}^{P}\tilde{B}_{ai}^{Q}$		OVX^2T
9	Make $\tilde{x}_{PQ}$, $\tilde{y}_{PQ}$, $\tilde{w}_{PQ}$, $\tilde{z}_{PQ}$, $\tilde{e}_{PQ}$, $\overline{f}^{P}$, $\overline{h}^{P}$		
10	Write $\tilde{x}_{PQ}$, $\tilde{y}_{PQ}$, $\tilde{w}_{PQ}$, $\tilde{z}_{PQ}$, $\tilde{e}_{PQ}$		
11	g^{P} += $\sum_t \rho_t \tilde{x}_{PQ}\tilde{f}^{Q}$		
12	γ_{RS} -= $c_T\rho_t\sum_Q \tilde{x}_{RQ}\tilde{x}_{SQ}$		
13	γ_{RS}^{Π} += $c_T\rho_t\sum_Q(\tilde{e}_{RQ}\tilde{x}_{SQ} + \tilde{x}_{RQ}\tilde{e}_{SQ}) - \overline{h}^{R}\overline{f}^{S}$		
14	γ_{RS}^{D} -= $c_U\rho_t\sum_Q(\tilde{z}_{RQ}\tilde{x}_{SQ} + \tilde{x}_{RQ}\tilde{z}_{SQ} + \tilde{y}_{RQ}\tilde{w}_{SQ} + \tilde{w}_{RQ}\tilde{y}_{SQ})$		

to an ill-behavedness of this factor, especially for low-energy core orbitals and high-energy virtuals and will seriously affect the evaluations of 1PDMs if we directly follow eqs 24, 25, 40, and 41. Thus, the 1PDM equations need to be modified to prevent this potential problem.

We can always avoid this issue if we only use $e^{-\varepsilon_a t}$ and $e^{\varepsilon_i t}$ in the working expressions, as the virtual and occupied energies tend to be positive and negative, respectively. Occasional sign switches will not pose any difficulties, as absolute values of such orbital energies will be rather small. Thus, by defining $\bar{\Lambda}^Q_{bj}$ as

$$\bar{\Lambda}^P_{ai} = \tilde{\Lambda}^P_{at} e^{\varepsilon_a t} e^{-\varepsilon_i t} = {}^{\beta}\tilde{e}_{PQ} B^Q_{ai} \tag{61}$$

and by defining $\bar{\Omega}^P_{ai}$, $\bar{\mathrm{M}}^P_{ai}$, and $\bar{\mathrm{N}}^P_{ai}$ analogously, we can rewrite the 1PDM expressions as follows

$$P^{\Pi}_{ab} = -\sum_t \rho_t \frac{e^{-\varepsilon_b t} - e^{-\varepsilon_a t}}{\Delta^a_b} (\bar{\Lambda}^Q_{bj} + \bar{\Omega}^Q_{bj})(aj|Q) e^{\varepsilon_j t} \tag{62}$$

$$P^{\Pi}_{ij} = \sum_t \rho_t \frac{e^{\varepsilon_j t} - e^{\varepsilon_i t}}{\Delta^j_i} (\bar{\Lambda}^Q_{bj} + \bar{\Omega}^Q_{bj})(ib|Q) e^{-\varepsilon_b t} \tag{63}$$

$$P^{\mathrm{D}}_{ab} = -\sum_t \rho_t \frac{e^{-\varepsilon_b t} - e^{-\varepsilon_a t}}{\Delta^a_b} (\bar{\mathrm{M}}^Q_{bj} A^Q_{aj} + \bar{\mathrm{N}}^Q_{bj}(aj|Q)) e^{\varepsilon_j t} \tag{64}$$

$$P^{\mathrm{D}}_{ij} = \sum_t \rho_t \frac{e^{\varepsilon_j t} - e^{\varepsilon_i t}}{\Delta^j_i} (\bar{\mathrm{M}}^Q_{bj} A^Q_{bi} + \bar{\mathrm{N}}^Q_{bj}(ib|Q)) e^{-\varepsilon_b t} \tag{65}$$

As is proven in the Appendix, these equations are still valid even with degenerate orbitals when the function $f(p,q) = (e^{pt} - e^{qt})/(p - q)$ is analytically continued toward its singular point ($p = q$). Practically, the following expansion should be adopted to preserve enough numerical precisions in near-degenerate cases

$$\frac{e^{pt} - e^{qt}}{p - q} = t e^{(p+q)t/2}\left(1 + \frac{1}{2^2 \cdot 3!}\delta^2 t^2 + \frac{1}{2^4 \cdot 5!}\delta^4 t^4 + \ldots\right) \tag{66}$$

with $\delta = p - q$. We have exclusively used these 1PDM

Chart 2. Algorithm of Laplace Loop II (Path 1) and Its CPU and IO Costs

		Disk IO cost	CPU cost
	Loop over t ($t \in$ quadrature points, size: T)		
1	Read $\tilde{x}_{PQ}$, $\tilde{y}_{PQ}$, $\tilde{w}_{PQ}$, $\tilde{z}_{PQ}$, $\tilde{e}_{PQ}$		
	Loop over i ($i \in$ OCC)		
2	Read B^P_{ai} , D^P_{ai} , $E^P_{ai} = W^P_{ai} - Q^P_{ai}$	$3OVXT$	
3	$H_{ai} \mathrel{+}= \rho_t \sum_P B^P_{ai} e^{-(\varepsilon_a - \varepsilon_i)t} \tilde{f}^P$		
4	$\bar{\Gamma}^P_{ai} = -\tilde{x}_{PQ} B^Q_{ai}$		OVX^2T
5	$\bar{\Omega}^P_{ai} = \tilde{x}_{PQ} E^Q_{ai} - b^a_i \bar{h}^P - I_{ai} \bar{f}^P$		OVX^2T
6	$\bar{\Lambda}^P_{ai} + r\bar{\mathrm{N}}^P_{ai} = \left(\tilde{e}_{PQ} - r\tilde{z}_{PQ}\right) B^Q_{ai} - r\tilde{w}_{PQ} D^Q_{ai}$		$2OVX^2T$
7	$r\bar{\mathrm{M}}^P_{ai} = -r\left(\tilde{x}_{PQ} D^Q_{ai} + \tilde{y}_{PQ} B^Q_{ai}\right)$		$2OVX^2T$
8	Save $\bar{\Gamma}^P_{ai}$	$OVXT$	
9	Save $\bar{\omega}^P_{ai} = \bar{\Gamma}^P_{ai} + \bar{\Omega}^P_{ai} + \bar{\Lambda}^P_{ai} + r\bar{\mathrm{N}}^P_{ai}$	$OVXT$	
10	Save $\bar{\lambda}^P_{ai} = \bar{\Omega}^P_{ai}/2 + \bar{\Lambda}^P_{ai} + r\bar{\mathrm{N}}^P_{ai}$	$OVXT$	
11	Save $\bar{\mu}^P_{ai} = r\bar{\mathrm{M}}^P_{ai}$	$OVXT$	
	Loop over i-batch ($i \in$ OCC)		
12	If ($t > 0$), read Γ^P_{ai} , λ^P_{ai} , μ^P_{ai}	$3OVX(T-1)$	
13	Read $\bar{\Gamma}^P_{ai}$, $\bar{\lambda}^P_{ai}$, $\bar{\mu}^P_{ai}$	$3OVXT$	
14	$\Gamma^P_{ai} \mathrel{+}= c_T \rho_t \bar{\Gamma}^P_{ai} e^{-\varepsilon_a t} e^{\varepsilon_i t}$		
15	$\lambda^P_{ai} \mathrel{+}= c_T \rho_t \bar{\lambda}^P_{ai} e^{-\varepsilon_a t} e^{\varepsilon_i t}$		
16	$\mu^P_{ai} \mathrel{+}= c_T \rho_t \bar{\mu}^P_{ai} e^{-\varepsilon_a t} e^{\varepsilon_i t}$		
17	Write Γ^P_{ai} , λ^P_{ai} , μ^P_{ai}	$3OVXT$	
18	Transpose $\bar{\omega}^P_{ai}$, $\bar{\mu}^P_{ai}$ on disk: $(a, i, P) \leftarrow (a, P, i)$	$4OVXT$	
	Loop over P ($P \in$ AUX)		
19	Read $\bar{\omega}^P_{ai}$, $\bar{\mu}^P_{ai}$	$2OVXT$	
20	Read $(a,i\,|\,P)$, A^P_{ai}	$2OVXT$	
21	$\tilde{P}_{ab} \mathrel{-}= \sum_i \bar{\omega}^P_{ai} (bi\,|\,P) e^{\varepsilon_i t}$		OV^2XT
22	$\tilde{P}_{ij} \mathrel{+}= \sum_a \bar{\omega}^P_{ai} (aj\,|\,P) e^{-\varepsilon_a t}$		O^2VXT
23	$\tilde{P}_{ab} \mathrel{-}= \sum_i \bar{\mu}^P_{ai} A^P_{bi} e^{\varepsilon_i t}$		OV^2XT
24	$\tilde{P}_{ij} \mathrel{+}= \sum_a \bar{\mu}^P_{ai} A^P_{aj} e^{-\varepsilon_a t}$		O^2VXT
25	$P^{\mathrm{corr}}_{ab} \mathrel{+}= c_T \rho_t \tilde{P}_{ab} \left(e^{-\varepsilon_b t} - e^{-\varepsilon_a t}\right)/\Delta^a_b$		
26	$P^{\mathrm{corr}}_{ij} \mathrel{+}= c_T \rho_t \tilde{P}_{ij} \left(e^{\varepsilon_j t} - e^{\varepsilon_i t}\right)/\Delta^j_i$		
27	$P^{\mathrm{corr}}_{ab} \leftarrow P^{\mathrm{corr}}_{ab} + P_{ab}$		
28	$P^{\mathrm{corr}}_{ij} \leftarrow P^{\mathrm{corr}}_{ij} + P_{ij}$		

Chart 3. Algorithm of Laplace Loop II (Path 2) and Its CPU and IO Costs

		Disk IO cost	CPU cost
1	Read $\tilde{x}_{PQ}$, $\tilde{y}_{PQ}$, $\tilde{w}_{PQ}$, $\tilde{z}_{PQ}$, $\tilde{e}_{PQ}$ for all quadrature points		
	Loop over i ($i \in$ OCC)		
2	Read B_{ai}^P , D_{ai}^P , $E_{ai}^P = W_{ai}^P - Q_{ai}^P$	$3OVX$	
	Loop over t ($t \in$ quadrature points, size: T)		
3	H_{ai} += $\rho_t \sum_P B_{ai}^P e^{-(\varepsilon_a - \varepsilon_i)t} \tilde{f}^P$		
4	$\bar{\Gamma}_{ai}^P = -\tilde{x}_{PQ} B_{ai}^Q$		OVX^2T
5	$\bar{\Omega}_{ai}^P = \tilde{x}_{PQ} E_{ai}^Q - b_i^a \bar{h}^P - I_{ai} \bar{f}^P$		OVX^2T
6	$\bar{\Lambda}_{ai}^P + r\bar{\mathrm{N}}_{ai}^P = \left(\tilde{e}_{PQ} - r\tilde{z}_{PQ}\right) B_{ai}^Q - r\tilde{w}_{PQ} D_{ai}^Q$		$2OVX^2T$
7	$\bar{\mathrm{M}}_{ai}^P = -\left(\tilde{x}_{PQ} D_{ai}^Q + \tilde{y}_{PQ} B_{ai}^Q\right)$		$2OVX^2T$
8	Γ_{ai}^P += $c_T \rho_t \bar{\Gamma}_{ai}^P e^{-\varepsilon_a t} e^{\varepsilon_i t}$		
9	λ_{ai}^P += $c_T \rho_t \left(\bar{\Omega}_{ai}^P / 2 + \bar{\Lambda}_{ai}^P + r\bar{\mathrm{N}}_{ai}^P\right) e^{-\varepsilon_a t} e^{\varepsilon_i t}$		
10	μ_{ai}^P += $c_U \rho_t \bar{\mathrm{M}}_{ai}^P e^{-\varepsilon_a t} e^{\varepsilon_i t}$		
11	Write $\bar{\omega}_{ai}^P(t) = \bar{\Gamma}_{ai}^P + \bar{\Omega}_{ai}^P + \bar{\Lambda}_{ai}^P + r\bar{\mathrm{N}}_{ai}^P$	$OVXT$	
12	Write $\bar{\mu}_{ai}^P(t) = r\bar{\mathrm{M}}_{ai}^P$	$OVXT$	
13	Write Γ_{ai}^P , λ_{ai}^P , μ_{ai}^P	$3OVX$	
14	Transpose $\bar{\omega}_{ai}^P$, $\bar{\mu}_{ai}^P$ on disk: $(a, t, i, P) \leftarrow (a, P, t, i)$	$4OVXT$	
	Loop over P ($P \in$ AUX)		
15	Read $(a,i \mid P)$, A_{ai}^P	$2OVX$	
	Loop over t ($t \in$ quadrature points, size: T)		
16	Read $\bar{\omega}_{ai}^P$, $\bar{\mu}_{ai}^P$	$2OVXT$	
17	$\tilde{P}_{ab} = \sum_t \bar{\omega}_{ai}^P (bi \mid P) e^{\varepsilon_i t} + \sum_t \bar{\mu}_{ai}^P A_{bi}^P e^{\varepsilon_i t}$		$2OV^2XT$
18	$\tilde{P}_{ij} = \sum_a \bar{\omega}_{aj}^P (aj \mid P) e^{-\varepsilon_a t} + \sum_a \bar{\mu}_{ai}^P A_{aj}^P e^{-\varepsilon_a t}$		$2O^2VXT$
19	P_{ab}^{corr} -= $c_T \rho_t \tilde{P}_{ab} \left(e^{-\varepsilon_b t} - e^{-\varepsilon_a t}\right) / \Delta_b^a$		
20	P_{ij}^{corr} += $c_T \rho_t \tilde{P}_{ij} \left(e^{\varepsilon_j t} - e^{\varepsilon_i t}\right) / \Delta_i^j$		
21	$P_{ab}^{\mathrm{corr}} \leftarrow P_{ab}^{\mathrm{corr}} + P_{ab}$		
22	$P_{ij}^{\mathrm{corr}} \leftarrow P_{ij}^{\mathrm{corr}} + P_{ij}$		

representations in the actual implementation, which will be presented in the next section.

III. Implementation

As one can see in the above section, SOS-CIS(D_0) gradient equation involves numerous intermediate terms. In addition, the excited-state gradient calculation should always include the concurrent evaluation of the ground-state gradient. Therefore, the routines that calculate these various intermediates must be carefully designed to minimize the processor and the disk usage. For this purpose, the gradient is obtained with the following algorithm.

A. Pre-Laplace Quadrature Loop. The purpose of this routine is to build intermediates that will be used in the main Laplace loop detailed in the upcoming subsections. All the quadrature-independent terms listed in Table 1 need to be evaluated here except H_{ai} and g^P. Besides the fact that their formulation is quite straightforward, its computational cost is rather small when it is compared to the cost of the Laplace loop, because this routine is executed only once. In addition, certain terms can be inherited from the result of the iteration for the excitation amplitude **b**. Therefore, we will not elaborate on the detailed performance consideration of this part.

B. Laplace Quadrature Loop I. This part aims to form various Laplace quadrature dependent intermediates and portions of 2C-2PDMs. As this routine involves a significant amount of fourth-order steps in each pass of the Laplace loop, its efficiency is an important issue for consideration. The detailed scheme is presented in Chart 1, together with its expected processor cost.

C. Laplace Quadrature Loop II: Path 1. This routine performs the main calculations for the various 3C-2PDMs and the 1PDMs that will be later used in the gradient evaluation through integrals. Unlike the ground-state SOS-MP2 analytical gradient case,[36] because the forms of 3C-2PDMs adopted in the 3-centered force evaluation (eq 19) and 1PDM generation (eqs 24 and 25) are somewhat different, the various components of 3C-2PDMs need to be stored on disk individually. This will significantly increase the disk IO cost compared to the equivalent portion of the fourth-order scaling ground-state gradient theory.[36] In Chart 2, one can see that the 3C-2PDMs are scaled and combined in varying forms ($\bar{\omega}_{ai}^P$, $\bar{\lambda}_{ai}^P$, and $\bar{\mu}_{ai}^P$). The purpose of using these altered forms of 2PDMs together with an intermediate scaling parameter, $r = c_U/c_T$, is to reduce this IO cost. After this modification, the associated IO cost is only about twice of the ground-state case.

D. Laplace Quadrature Loop II: Path 2. In fact, the IO cost of Path 1 presented above can be further improved when we are willing to pay more cost for memory and disk storages. This can easily be inferred by observing that much of the IO cost in Path 1 is involved with the increment of 2PDMs at different Laplace quadrature points (lines 12−17

Chart 4. Algorithm of Post-Laplace Loop and Its CPU and IO Costs

		Disk IO cost	CPU cost
	Loop over i ($i \in$ OCC)		
1	Read B_{ai}^{P}, $V_{j(i)}^{P}$	ONX	
2	$v_{j(i)}^{P} = \sum_{Q} V_{j(i)}^{Q} (Q \mid P)^{-1/2}$		O^2X^2
3	${}_1\mathrm{X}_{ai}^{P} = \sum_{j} H_{aj} v_{j(i)}^{P}$		O^2VX
4	Write ${}_1\mathrm{X}_{ai}^{P}$	OVX	
5	$C_{ai}^{P} = \sum_{Q} B_{ai}^{Q} (Q \mid P)^{-1/2}$		OVX^2
6	$w_{(j)i}^{P} = \sum_{a} H_{aj} C_{ai}^{P}$		O^2VX
7	Write $w_{(j)i}^{P}$	O^2X	
8	$\gamma_{RS}^{\Pi} \mathrel{+}= c_T \sum_{j} v_{j(i)}^{R} w_{(j)i}^{S}$		
9	$\gamma_{RS}^{\Pi} \mathrel{-}= c_T g^{R} \sum_{P} ({}^{\alpha}f^{P} + {}^{\beta}f^{P})(P \mid R)^{-1/2}$		
10	Increment λ_{ai}^{P} on disk: $\lambda_{ai}^{P} \leftarrow \lambda_{ai}^{P} + c_T \cdot {}_1\mathrm{X}_{ai}^{P}/2$	$3OVX$	
11	Transpose $w_{(i),j}^{P}$ on disk: $(P, (i), j) \leftarrow (P, j, (i))$	$2O^2X$	
	Loop over i ($i \in$ OCC)		
12	Read $w_{j(i)}^{P}$	O^2X	
13	${}_2\mathrm{X}_{ai}^{P} = \sum_{j} b_{j}^{a} w_{j(i)}^{P}$		
14	${}_2\mathrm{X}_{ai}^{P} \mathrel{-}= H_{ai} \sum_{R} ({}^{\alpha}f^{R} + {}^{\beta}f^{R})(P \mid R)^{-1/2}$		
15	${}_2\mathrm{X}_{ai}^{P} \mathrel{-}= b_{i}^{a} ({}^{\alpha}g^{P} + {}^{\beta}g^{P})$		
16	Write ${}_2\mathrm{X}_{ai}^{P}$	OVX	
17	Increment λ_{ai}^{P} on disk: $\lambda_{ai}^{P} \leftarrow \lambda_{ai}^{P} + c_T \cdot {}_2\mathrm{X}_{ai}^{P}/2$	$3OVX$	
18	Transpose λ_{ai}^{P} on disk: $(a, i, P) \leftarrow (a, P, i)$	$2OVX$	
19	Transpose Γ_{ai}^{P} on disk: $(a, i, P) \leftarrow (a, P, i)$	$2OVX$	
	Loop over P ($P \in$ AUX)		
20	Read Γ_{ai}^{P}, λ_{ai}^{P}	$2OVX$	
21	${}_3\mathrm{X}_{ai}^{P} = -\sum_{b} P_{ab} \Gamma_{bi}^{P} + \sum_{j} P_{ij} \Gamma_{aj}^{P}$		$OVNX$
22	Write $\lambda_{ai}^{P} \leftarrow \lambda_{ai}^{P} + \Gamma_{ai}^{P} + {}_3\mathrm{X}_{ai}^{P}/2$	OVX	

in Chart 2). At the first glance, one might expect that this cost can be significantly reduced if the 2PDMs are kept in the core memory. However, this will require cubic memory storage, which will be unacceptably large for many sizable molecules.

The IO cost reduction can actually be attained by switching the order of Laplace and occupied loops. Namely, when the Laplace loop is executed for one selected occupied orbital, 2PDM increments can be performed within the core memory while keeping the memory requirement to be quadratic. This implementation is shown in Chart 3 along with its associated IO and CPU cost. From Charts 2 and 3, one can deduce that the IO cost reduction from this algorithm is from $24OVXT$ down to $\sim 9OVXT$ (assuming $T = 7$[30,32]).

There are two drawbacks for adopting this algorithm. These are the necessity for storing two quadrature-dependent densities ($\bar{\omega}_{ai}^{P}$ and $\bar{\mu}_{ai}^{P}$) on disk at all quadrature points, and the use of T-times larger memory for storing $\tilde{x}$, $\tilde{y}$, $\tilde{z}$, $\tilde{w}$, and $\tilde{e}$ (line 1 in Chart 3). However, the disk storage is usually abundant in modern computational platforms, and this shortcoming associated with the necessary disk space will be mostly insignificant in any practical calculations. The memory requirement will be similarly insignificant: even for a system with 10,000 auxiliary basis functions, the required space for storing the above five matrices will only be 4 gigabytes. More importantly, the required disk and memory storage can always be calculated before entering the Laplace Loop II, and when the available space becomes an issue, Path 1 can be followed as a fallback mechanism without any difficulties.

E. Post-Laplace Quadrature Loop for 2PDM. The remaining 2PDM components after completing the Laplace loops are evaluated in this routine. Even though there are various elements computed in this routine as detailed in Chart 4, the associated computational cost is relatively small as this routine does not involve the quadrature points. After completing this routine, all 2PDMs are accumulated and stored on disk for their subsequent use in the evaluation of the force terms through the use of integrals in the atomic orbital basis.[57]

F. Force and Lagrangian Evaluations Using Atomic Orbital Based Integrals. With all the density matrices prepared in the above, we are now ready to compute various force terms using the integrals with atomic orbital basis. This is performed concurrently with the construction of Lagrangians in the mixed representations (eqs 44 and 45) as it uses the same density matrices and their transforms[44] as shown in Chart 5. Practically, this part contains two calls of one identical subroutine with two different AO to MO transformation coefficients ($C_{\mu i}$ and $D_{\mu i}$) and two different sets of 2PDM matrices (λ_{ai}^{P} and μ_{ai}^{P} as defined in Charts 2 and 3). The mathematical background for doing this is explained in the previous section. As the structure of the common subroutine closely resembles the one in RI-MP2 gradient calculation for obtaining **L** and 3-centered force,[44] this part can be easily implemented with minor modifications on any

Chart 5. Algorithm of Force and Lagrangian Evaluations through Atomic Orbital Based Integrals

		Disk IO cost	CPU cost
	Loop over P ($P \in$ AUX)		
1	Read λ_{ai}^{P}, μ_{ai}^{P}	$2OVX$	
2	Back-transform: $\lambda_{\mu i}^{P} = \sum_a C_{\mu a}\lambda_{ai}^{P}$	ONX	$OVNX$
3	Back-transform: $\mu'^{P}_{\mu i} = \sum_{a,j}\left(C_{\mu j}\mu_{aj}^{P}b_i^a + C_{\mu j}\mu_{ai}^{P}b_j^a\right)$	ONX	$2O^2(V+N)X$
	Loop over P-batch ($P \in$ AUX)		
4	Read λ_{ai}^{P}, $\lambda_{\mu i}^{P}$, $\mu'^{P}_{\mu i}$	$\sim 3ONX$	
5	Make $(\mu\nu \mid P)$		
6	$L_{vi}^{\mathrm{I},2} \mathrel{+}= \sum_{\mu,P}\left(\lambda_{\mu i}^{P} - \mu'^{P}_{\mu i}\right)(\mu\nu \mid P)$		ON^2X
7	Transform: $(i\nu \mid P) = \sum_\mu C_{i\mu}(\mu\nu \mid P)$		ON^2X
8	$L_{av}^{\mathrm{I},1} \mathrel{+}= \sum_{i,P}\lambda_{ai}^{P}(i\nu \mid P)$		$OVNX$
9	Make $(\mu\nu \mid P)^x$		
10	$E^{(x)} \mathrel{+}= \sum_{i,\mu,\nu,P}\left(2\lambda_{\mu i}^{P} - \mu'^{P}_{\mu i}\right)C_{vi}(\mu\nu \mid P)^x$		$\sim ON^2X$
	Loop over P ($P \in$ AUX)		
11	Read μ_{ai}^{P}	OVX	
12	Back-transform: $\mu_{\mu i}^{P} = \sum_a C_{\mu a}\mu_{ai}^{P}$	ONX	$OVNX$
	Loop over P-batch ($P \in$ AUX)		
13	Read μ_{ai}^{P}, $\mu_{\mu i}^{P}$	$\sim 2OVX$	
14	Back-transform: $\mu_{\mu i}^{P} = \sum_a C_{\mu a}\mu_{ai}^{P}$	ONX	$OVNX$
15	Make $(\mu\nu \mid P)$		
16	$L_{vi}^{\mathrm{II},2} \mathrel{+}= \sum_{\mu,P}\mu'^{P}_{\mu i}(\mu\nu \mid P)$		ON^2X
17	Transform: $(i\nu \mid P)' = \sum_\mu D_{i\mu}(\mu\nu \mid P)$		ON^2X
18	$L_{av}^{\mathrm{II},1} \mathrel{+}= \sum_{i,P}\mu_{ai}^{P}(i\nu \mid P)'$		$OVNX$
19	$E^{(x)} \mathrel{+}= 2\sum_{i,\mu,\nu,P}\mu_{\mu i}^{P}D_{vi}(\mu\nu \mid P)^x$		$\sim ON^2X$
20	$L_{av}^{1} = L_{av}^{\mathrm{I},1} + L_{av}^{\mathrm{II},1} + \sum_i L_{vi}^{\mathrm{II},2}b_i^a$, $L_{vi}^{2} = L_{vi}^{\mathrm{I},2}$		
21	Make $(P \mid Q)^x$		
22	$E^{(x)} \mathrel{-}= \sum_{P,Q}\gamma_{PQ}^{\mathrm{total}}(P \mid Q)^x$		

Table 2. CPU Times for Calculating SOS-MP2 Ground-State and SOS-CIS(D_0) Excited-State Gradients[a]

molecule	no. of basis	CPU time (min)		ratio
		SOS-MP2	SOS-CIS(D_0)	
hexatriene (C_6H_8)	460[b]	139 (88)	181 (119)	1.30 (1.35)
styrene (C_8H_8)	552[b]	341 (202)	462 (283)	1.35 (1.40)
azulene($c_{10}H_8$)	644[b]	691 (347)	884 (469)	1.28 (1.35)
anthracene ($C_{14}H_{10}$)	874[b]	1805 (1031)	2002 (1290)	1.11 (1.25)
pyrene ($C_{16}H_{10}$)	966[b]	3183 (1851)	3090 (1988)	0.97 (1.07)
ZnBC-BC ($C_{46}H_{36}N_8Zn$)[c]	918[d]	576 (450)	1667 (1510)	2.90 (3.36)

[a] Measured with a 2.0 GHz Opteron processor. Numbers in parentheses are based on CPU times excluding the load for solving **Z**-vectors. Timing data exclude the cost for solving the HF-SCF equation and the single excitation amplitude **b**. [b] With aug-cc-pVTZ and its corresponding auxiliary basis sets (refs 58 and 59). [c] Structure given in ref 38. [d] With 6-31G(d) and the VDZ auxiliary basis sets (ref 15).

existing RI-MP2 gradient code. This practical convenience is why there are two separate integration batching loops in Chart 5.

IV. Demonstration and Concluding Remarks

The above analytical SOS-CIS(D_0) gradient algorithm has been implemented in the development version of Q-Chem 3.1,[57] and its performance was benchmarked with selected molecules with varying sizes. The molecules adopted during this benchmark process are listed in Table 2, together with the measured computational costs. Since we aim our new method to have similar efficiency as its sister ground-state gradient theory SOS-MP2,[36] comparing the performances of the two methods will be an important task. Thus, in Table 2, we have presented the timing data from SOS-MP2 gradient calculation as well as the timing for our new SOS-CIS(D_0) gradient. For both methods, as the CPU time has been found to be very close to the wall-clock time in most cases, we will only focus on comparing the CPU time.

From Table 2, we can clearly see that the efficiency of our SOS-CIS(D_0) gradient is very close to that of SOS-MP2 gradient. In fact, a detailed comparison of the algorithm charts presented in the previous section with the ones in ref 36 reveals that the routine for calculating various 3-centered density matrices is $\sim 5\times$ more expensive than the corresponding part in SOS-MP2 theory, while performing 1PDM calculations, AO-based integrals, and their associated matrix transformations is approximately twice more expen-

Table 3. CPU Times for Calculating Analytical Gradients at Various Levels of Theory[a]

molecule	no. of basis	CPU time (min)		
		HF	CIS	SOS-CIS(D_0)
hexatriene (C_6H_8)	460[b]	0.3	134	181
styrene (C_8H_8)	552[b]	0.5	339	462
azulene ($C_{10}H_8$)	644[b]	1	730	884
anthracene ($C_{14}H_{10}$)	874[b]	3	1629	2002
pyrene ($C_{16}H_{10}$)	966[b]	4	2707	3090
ZnBC-BC ($C_{46}H_{36}N_8Zn$)	918[c]	44	182	1667

[a] Measured with a 2.0 GHz Opteron processor. Timing data exclude the cost for solving the HF-SCF equation and the single excitation amplitude **b**. [b] With aug-cc-pVTZ and its corresponding auxiliary basis sets (refs 58 and 59). [c] With 6-31G(d) and the VDZ auxiliary basis sets (ref 15).

sive. The common routine for building portions of **L** and **W** by contracting unrelaxed and relaxed 1PDM with AO-based ERIs will cost the same. The routine for solving the **Z**-vector equation will also cost similarly, although the actual relative cost will vary depending on the number of iterations required to reach convergences in the ground-state and excited-state cases. The result from Table 2 can be understood on this ground: the relative CPU cost is close to 1.0 when the load for obtaining the **Z**-vector solution is dominant (the case with the aug-cc-pVTZ basis[58]), while it reaches ~3.0 as other routines become more important. As explained above, our method will cost ~5× more even in the worst case scenario, and in any practical calculations we expect it to perform close to the efficient SOS-MP2 gradient.[36]

We can see a similar behavior when we compare the cost for obtaining CIS and SOS-CIS(D_0) gradients as presented in Table 3. Unless the system size becomes large enough as in the case of ZnBC-BC complex, the additional cost for including electron correlation effect (CIS → SOS-CIS(D_0)) in the gradient evaluation is effectively insignificant because the cost for executing the common routines mentioned in the above are dominant. Of course, evaluating HF gradient after SCF convergence does not require such routines (for example, HF gradient does not require the **Z**-vector), and its cost is almost negligible as shown in Table 3.

In fact, because the computation of the gradient always require executing pregradient routines for reaching HF-SCF solution and for obtaining **b**-amplitudes as eigensolutions of the response matrices, and because these pregradient routines are relatively expensive due to their iterative nature, the most practical comparison should be made by examining the total CPU costs associated with both pregradient and gradient routines. This comparison is schematically presented in Figure 1 for the selected molecules. Again, from this figure, one can be assured that the efficiency of SOS-CIS(D_0) is comparable to other presented methods in general. The only exception is the ZnBC-BC complex. In this molecule, AO-based HF-SCF and CIS solutions are obtained relatively fast because many function pairs are ignored in the integrals due to the large size the system and the relatively low quality of the adopted basis set.

In summary, we have shown that an analytic gradient of SOS-CIS(D_0) can be obtained with a fourth-order computational effort as in the case of its sister ground-state theory,

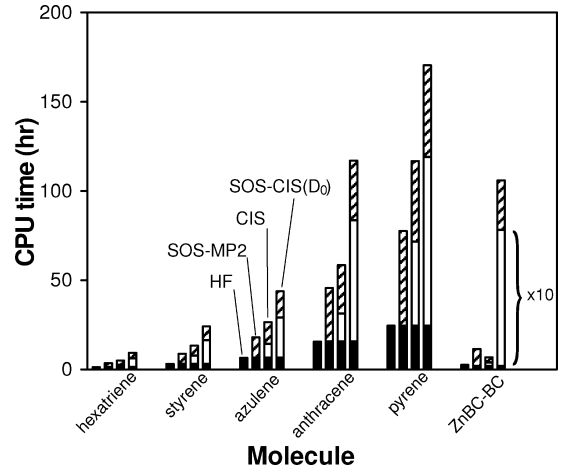


Figure 1. Analysis of total CPU times for obtaining analytical gradients at various levels of theory. Four vertical bars, from left to right, for each molecule represent the total CPU time with HF, SOS-MP2, CIS, and SOS-CIS(D_0) respectively, as denoted in the figure. Different segments of the bars are coded as follows: filled bars (HF-SCF timing), empty bars (CIS or SOS-CIS(D_0) excitation amplitude solution), and hashed bars (gradient calculation). The CPU time for reaching SOS-CIS(D_0) amplitude solution with ZnBC-BC was downscaled by a factor of 10 for visual clarity. A threshold of 10^{-12} au was employed for integral screening, while the convergence criteria for SCF and **b**-amplitude solutions were set to 10^{-7} au and 10^{-6} au, respectively. Other computational details are described in Table 3.

SOS-MP2. This efficiency is attained from the use of Laplace transform in combination with the RI approximation. After implementing this new theory into an efficient algorithm, the actual processor time for calculating the excited-state gradient was found to be comparable to the ground-state gradient calculation time for many molecules. In fact, because the SOS-CIS(D_0) excitation amplitude and energy evaluation involves about 10 additional iteration steps after the conventional SCF solution, the cost for evaluating gradient will be effectively minor. This was indeed true at least for the systems adopted during our benchmark with at most 37% of post SCF calculation time spent on the gradient calculation.

Accordingly, as an efficient wave function-based excited-state gradient theory with a capability of properly describing quasidegenerate states, this method will be useful in studying various quantum chemical systems. Its utility will be diverse ranging from performing a simple excited-state geometry optimization to studying complex dynamics in various photophysical systems. Indeed, we note that various excited-state geometry searches have relied on the electron-uncorrelated CIS gradient, even when the resulting geometries were used for parametrizing an electron-correlated excited theory such as SOS-CIS(D).[30,32] This was simply due to the previous lack of an efficient wave function-based electron-correlated gradient theory. Thus, applying the present method to such parametrization processes along with numerical comparisons of the various excited-state parameters will be an interesting and important task. We hope to report on this development in due course.

Acknowledgment. This work was supported by a subcontract from Q-Chem Inc, from an NIH SBIR grant, and was also supported by the Office of Basic Energy Sciences of the U.S. Department of Energy through the LBL Ultrafast Center. We are grateful for a grant of supercomputer time from NERSC. D.C. acknowledges financial support from a Fulbright Fellowship. M.H.G. is a part-owner of Q-Chem Inc.

Appendix A: Derivation of $[e^{-\Delta t}]^{(x)}$

From $e^{-\Delta t} = \sum_{n=0}^{\infty}(-)^n\Delta^n t^n/n!$, we can express the derivative as

$$[e^{-\Delta t}]^{(x)} = \sum_{n=1}^{\infty}\sum_{m=0}^{n-1}\frac{1}{n!}(-)^n\Delta^{n-m-1}\Delta^{(x)}\Delta^m t^n$$

$$= \sum_{n=0}^{\infty}\sum_{m=0}^{n}\frac{1}{(n+1)!}(-)^{n+1}\Delta^{n-m}\Delta^{(x)}\Delta^m t^{n+1} \tag{A1}$$

Note that we can always set Δ as diagonal after SCF completes

$$\Delta_{ijab;i'j'a'b'} = \delta_{ii'}\delta_{aa'}\delta_{jj'}\delta_{bb'}(\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j) = \Delta_{ij}^{ab} \tag{A2}$$

and $\Delta^{(x)}$ is block-diagonal as

$$\Delta^{(x)}_{ijab;i'j'a'b'} = \delta_{ii'}\varepsilon^{(x)}_{aa'}\delta_{jj'}\delta_{bb'} + \delta_{ii'}\delta_{aa'}\delta_{jj'}\varepsilon^{(x)}_{bb'} - \varepsilon^{(x)}_{ii'}\delta_{aa'}\delta_{jj'}\delta_{bb'} - \delta_{ii'}\delta_{aa'}\varepsilon^{(x)}_{jj'}\delta_{bb'} \tag{A3}$$

Because of this diagonal and block-diagonal nature, the matrix multiplications in eq A1 can be performed with relative ease, and its elements can be expressed as

$$[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'} = \sum_{n=0}^{\infty}\sum_{m=0}^{n}\frac{1}{(n+1)!}(-)^{n+1}t^{n+1}(\Delta_{ij}^{ab})^{n-m}\times [(\Delta_{ij}^{a'b})^m\varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'} + (\Delta_{ij}^{ab'})^m\delta_{ii'}\delta_{aa'}\delta_{jj'}\varepsilon^{(x)}_{bb'} - (\Delta_{i'j}^{ab})^m\varepsilon^{(x)}_{ii'}\delta_{aa'}\delta_{jj'}\delta_{bb'} - (\Delta_{ij'}^{ab})^m\delta_{ii'}\delta_{aa'}\varepsilon^{(x)}_{jj'}\delta_{bb'}] \tag{A4}$$

Because $\Delta_{ij}^{a'b} = \Delta_{ij}^{ab} + \Delta_a^{a'}$ with $\Delta_a^{a'} = \varepsilon_{a'} - \varepsilon_a$, the first term in eq A4 becomes

$$^I[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'} = \sum_{n=0}^{\infty}\sum_{m=0}^{n}\frac{1}{(n+1)!}(-)^{n+1}t^{n+1}\times (\Delta_{ij}^{ab})^{n-m}(\Delta_{ij}^{ab} + \Delta_a^{a'})^m\varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'}$$

$$= \sum_{n=0}^{\infty}\sum_{m=0}^{n}\frac{1}{(n+1)!}(-)^{n+1}t^{n+1}(\Delta_{ij}^{ab})^{n-m}\times \sum_{r=0}^{m}{}_mC_r(\Delta_a^{a'})^r(\Delta_{ij}^{ab})^{m-r}\varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'} \tag{A5}$$

with the use of a binomial expansion. Now, we can rearrange the summations to simplify this expression. From the following two relations

$$\sum_{n=0}^{\infty}\sum_{m=0}^{n}f(m,n) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}f(m,n+m) \tag{A6}$$

$$\sum_{m,n=0}^{\infty}f(m,m+n) = \sum_{s=0}^{\infty}\sum_{m=0}^{s}f(m,s) \tag{A7}$$

we can show that eq A5 can be transformed as

$$^I[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'} = \varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\sum_{m=0}^{s}\frac{1}{(s+r+1)!}(-t)^{s+r+1}\times (\Delta_{ij}^{ab})^s(\Delta_a^{a'})^r{}_{m+r}C_m$$

$$= \varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{1}{(s+r+1)!}(-t)^{s+r+1}\times (\Delta_{ij}^{ab})^s(\Delta_a^{a'})^r\sum_{m=0}^{s}{}_{m+r}C_m \tag{A8}$$

One can also easily prove that

$$\sum_{m=0}^{s}{}_{m+r}C_m = {}_{r+s+1}C_s \tag{A9}$$

is satisfied, and when $\Delta_a^{a'} \neq 0$, eq A8 reduces to the following simple expression

$$^I[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'} = \varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb}e^{-\Delta_{ij}^{ab}t}\frac{1}{\Delta_a^{a'}}[e^{-\Delta_a^{a'}t} - 1] \tag{A10}$$

If $\Delta_a^{a'} = 0$, the binomial expansion in eq A5 leads to

$$^I[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'}|_{\Delta_a^{a'}=0} = -t\varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb}e^{-\Delta_{ij}^{ab}t} \tag{A11}$$

and the analytic continuation of eq A10 to the pole of $\Delta_a^{a'}$ is a valid approach

$$^I[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'}|_{\Delta_a^{a'}=0} = \lim_{\Delta_a^{a'}\to 0}(^I[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'}) \tag{A12}$$

Namely, eq A10 can be used for any energy eigenvalues. Practically, when the energy difference $(\Delta_a^{a'})$ is smaller than some cutoff value, the Taylor expansion of $(1 - e^{-x})/x = 1 - x/2 + x^2/3! - x^3/4! + \cdots$ should be adopted to preserve the numerical precision.

The remaining three terms in eq A4 can be obtained in a similar manner, and the final expression for the matrix derivative is

$$[e^{-\Delta t}]^{(x)}_{ijab;i'j'a'b'} = -t\varepsilon^{(x)}_{aa'}\delta_{ii'}\delta_{jj'}\delta_{bb'}e^{-\Delta_{ij}^{ab}t}h(\Delta_a^{a'}t) - t\varepsilon^{(x)}_{bb'}\delta_{ii'}\delta_{jj'}\delta_{aa'}e^{-\Delta_{ij}^{ab}t}h(\Delta_b^{b'}t) + t\varepsilon^{(x)}_{ii'}\delta_{jj'}\delta_{aa'}\delta_{bb'}e^{-\Delta_{ij}^{ab}t}h(\Delta_i^{i}t) + t\varepsilon^{(x)}_{jj'}\delta_{ii'}\delta_{aa'}\delta_{bb'}e^{-\Delta_{ij}^{ab}t}h(\Delta_j^{j}t) \tag{A13}$$

with $h(x) = (1 - e^{-x})/x$.

References

(1) del Bene, J. E.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1971**, *55*, 2236.

(2) Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333.

(3) Monkhorst, H. J. *Int. J. Quantum Chem. Symp.* **1977**, *11*, 421.

(4) Mukherjee, D.; Mukherjee, P. *Chem. Phys.* **1979**, *39*, 325.

(5) Emrich, K. *Nucl. Phys. A* **1981**, *351*, 392.

(6) Sekino, H.; Bartlett, R. J. *Int. J. Quantum Chem. Symp.* **1984**, *18*, 225.

(7) Bartlett, R. J. *Modern Electronic Structure Theory*; World Scientific: Singapore, 1995.

(8) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.

(9) Gross, E. K. U.; Kohn, W. *Phys. Rev. Lett.* **1985**, *55*, 2850.

(10) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359.

(11) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514.

(12) Hättig, C.; Hald, K. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2111.

(13) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.

(14) Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97*, 331.

(15) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.

(16) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149.

(17) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496.

(18) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 3396.

(19) Almlöf, J. *Chem. Phys. Lett.* **1991**, *181*, 319.

(20) Häser, M.; Almlöf, J. *J. Chem. Phys.* **1992**, *96*, 489.

(21) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

(22) Ayala, P. Y.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 3660.

(23) Ayala, P. Y.; Kudin, K. N.; Scuseria, G. E. *J. Chem. Phys.* **2001**, *115*, 9698.

(24) Kats, D.; Usvyat, D.; Schütz, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3430.

(25) Izmaylov, A. F.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3421.

(26) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(27) Grimme, S.; Izgorodina, E. I. *Chem. Phys.* **2004**, *305*, 223.

(28) Szabados, A. *J. Chem. Phys.* **2006**, *125*, 214105.

(29) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.

(30) Rhee, Y. M.; Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 5314.

(31) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. *Chem. Phys. Lett.* **1994**, *219*, 21.

(32) Casanova, D.; Rhee, Y. M.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 164106.

(33) Head-Gordon, M.; Oumi, M.; Maurice, D. *Mol. Phys.* **1999**, *96*, 593.

(34) Hellweg, A.; Grun, S. A.; Hättig, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4119.

(35) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.

(36) Lochan, R. C.; Shao, Y. H.; Head-Gordon, M. *J. Chem. Theory Comput.* **2007**, *3*, 988.

(37) Casida, M. E.; Gutierrez, F.; Guan, J.; Gadea, F.-X.; Salahub, D.; Daudey, J.-P. *J. Chem. Phys.* **2000**, *113*, 7062.

(38) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007.

(39) Neugebauer, J.; Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2006**, *124*, 214101.

(40) Hieringer, W.; Görling, A. G. *Chem. Phys. Lett.* **2006**, *419*, 557.

(41) Dreuw, A.; Head-Gordon, M. *Chem. Phys. Lett.* **2006**, *426*, 231.

(42) Hieringer, W.; Görling, A. G. *Chem. Phys. Lett.* **2006**, *426*, 234.

(43) Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *129*, 034101.

(44) DiStasio Jr, R. A.; Steele, R. P.; Rhee, Y. M.; Shao, Y.; Head-Gordon, M. *J. Comput. Chem.* **2007**, *28*, 839.

(45) Maurice, D.; Head-Gordon, M. *Mol. Phys.* **1999**, *96*, 1533.

(46) Foresman, J. B.; Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *J. Phys. Chem.* **1992**, *96*, 135.

(47) Aikens, C. M.; Webb, S. P.; Bell, R. L.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. *Theor. Chem. Acc.* **2003**, *110*, 233.

(48) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; McGraw-Hill: New York, 1989.

(49) Pulay, P. *Mol. Phys.* **1969**, *17*, 197.

(50) Handy, N. C.; Amos, R. D.; Gaw, J. F.; Rice, J. E.; Simandrias, E. D. *Chem. Phys. Lett.* **1985**, *120*, 151.

(51) Head-Gordon, M. *Mol. Phys.* **1999**, *96*, 673.

(52) Rhee, Y. M.; DiStasio, R. A., Jr.; Lochan, R. C.; Head-Gordon, M. *Chem. Phys. Lett.* **2006**, *426*, 197.

(53) Handy, N. C.; Schaefer, H. F. *J. Chem. Phys.* **1984**, *81*, 503.

(54) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 281.

(55) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 275.

(56) Yamaguchi, Y.; Goddard, J. D.; Osamura, Y.; Schaefer , H. F., III *A New Dimension to Quantum Chemistry: Analytic Derivative Methods in Ab Initio Molecular Electronic Structure Theory*; Oxford University Press: New York, 1994.

(57) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; Distasio, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F., III; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(58) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(59) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175.

CT800509Z

JCTC Journal of Chemical Theory and Computation

Conformational Stability and Spin States of Cobalt(II) Acetylacetonate: CASPT2 and DFT Study

Mariusz Radoń,*,† Monika Srebro,† and Ewa Broclawik‡

Department of Theoretical Chemistry, Jagiellonian University, ul. Ingardena 3, 30-060 Kraków, Poland, and Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, ul. Niezapominajek 8, 30-239 Kraków, Poland

Received December 22, 2008

Abstract: Electronic structure and conformation of bis(acetylacetonate) cobalt(II), $Co(acac)_2$, a prototypical mediator in controlled radical polymerization of olefins, is reinvestigated. The ab initio multiconfigurational CASSCF/CASPT2 method is used to resolve the doubts stemming from density functional theory results. We assign the quartet ground state for a single molecule and point at tetrahedral conformation as the preferred one. Several density functionals are tested against the ab initio calculations, and their performance is assessed. The strength of intermolecular interactions in the crystal structure composed of square-planar $Co(acac)_2$ molecules (Burgess, J.; et al. *Acta Crystallogr.* 2000, *C56*, 649−650) is estimated to be sufficient for their planarization (suggested by Matyjaszewski, K.; et al. *Chem.−Eur. J.* 2007, *13*, 2480−2492).

1. Introduction

Bis(acetylacetonate) cobalt(II), $Co(acac)_2$, and its derivatives were recently proposed as mediators in controlled radical polymerization (CRP) of olefins. Cobalt-mediated CRP is based on the propensity of four-coordinated cobalt(II) complexes to reversibly capture organic radicals, thereby buffering their concentration at the desired level.[1] In theoretical studies on such reactions, the issues of great importance are to identify the spin symmetry and the lowest-energy conformation of the cobalt moiety and to investigate how these properties are affected by a coordination of an extra ligand (an organic radical). However, despite several studies treating $Co(acac)_2$, the issues of spin state and preferred conformation of this deceptively simple compound are not yet definitely closed.[2−7]

Two conformations were proposed for the $Co(acac)_2$ molecule: tetrahedral (of D_{2d} symmetry) and square-planar (of D_{2h} symmetry)[2−5] (Figure 1). According to the experiments from the 1960s, the $Co(acac)_2$ monomer was claimed to be tetrahedral,[3] which was also supported by a crystal structure resolution.[4] However, this crystal structure consists not of isolated molecules but of tetramers, where $Co(acac)_2$ units are chemically bound (share oxygens) and the coordination is effectively octahedral; therefore, this structure is not necessarily related to the isolated molecule. Quite surprisingly, a more recent (from 2000) crystal structure is built of square-planar $Co(acac)_2$ molecules, while no significant intermolecular interactions were claimed.[5] This recalled the original notion of square-planar conformation[2] (undermined in ref 3). The existence of stacking interaction between parallely arranged monomers in the crystal structure from ref 5 was suggested to affect the preferred conformation,[6] though the strength of these forces was not assessed at a quantitative level.

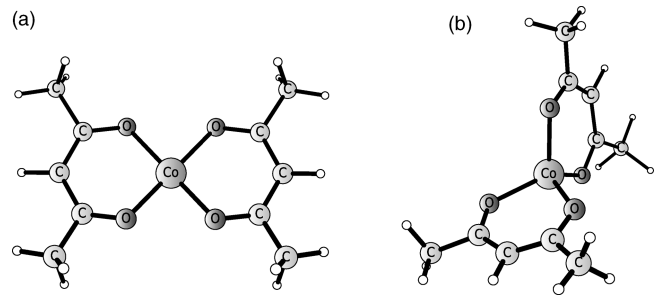


Figure 1. Alternative conformations for $Co(acac)_2$: (a) square-planar and (b) tetrahedral.

* Corresponding author e-mail: mradon@chemia.uj.edu.pl.
† Jagiellonian University.
‡ Polish Academy of Sciences.

10.1021/ct800571y CCC: $40.75
Published on Web 04/10/2009

In addition to two alternative conformations, two different spin states are possible for seven 3d electrons of cobalt(II): doublet (low-spin) and quartet (high-spin). No definite experimental evidence for any of them has been presented in the literature; nevertheless, the magnetic moment measurements from the 1960s opt for a high-spin ground state.[2] We note that some experiments on $Co(acac)_2$ are difficult to interpret or even inconclusive in terms of its electronic structure. For instance, the toluene solution of $Co(acac)_2$ has been described as EPR silent,[6,8] which is contradictory to a paramagnetic nature of both spin states, and caused presumably by a fast spin relaxation.

Recent density functional theory (DFT) calculations on $Co(acac)_2$ actually predicted the high-spin ground state in tetrahedral conformation,[6,9,10] yet these studies were based solely on the hybrid (B3LYP, B3PW91*) functionals, which are known to (sometimes erroneously) favor the high-spin states.[11] We shall demonstrate here that with a different choice of density functional quite opposite results can be found, what questions the DFT quality for the present problem and calls for ab initio calculations to resolve the doubts.

In this study, complete active space self-consistent field/second-order perturbation theory (CASSCF/CASPT2)[12,13] calculations are performed to unambiguously identify the ground state. This method is able to properly treat both exchange effects and strong (nondynamical) correlation present in transition-metal compounds,[14] what was proven in several successful applications, notably in two recent works for a variety of iron(II) complexes.[15,16] In the present study, CASPT2 energetics is contrasted with a DFT one and conclusions about the performance of various functionals are drawn.

As will be shown, high level ab initio calculations indicate a tetrahedral ground-state conformation, with strong support for the high-spin state. To shed more light on the reported crystal structural build of monomeric, square-planar $Co(acac)_2$ units,[5] thus at first sight contradicting our results, we also focus on prospective intermolecular interactions presumably causing a stacking arrangement.[6] To this end, we estimate intermolecular interaction energy in a dimer of two planar molecules (roughly modeling the crystal), to judge whether this interaction is sufficiently strong to impose the planarity of the $Co(acac)_2$ unit.

2. Methods and Models

The orientation of the $Co(acac)_2$ molecule used in this study is given in Figure 2, where also the twist angle θ between the acac rings is defined (acac = acetylacetonate).

The symmetry was employed in all calculations: D_{2h} for square-planar conformation, D_{2d} for tetrahedral conformation, and D_2 for a general (twisted) form. The symmetry labels of orbitals and electronic states are given below in D_2, as the highest common subgroup of all three groups.

For each conformation and electronic state, we performed (unrestricted) DFT geometry optimization (followed by frequency calculation and stability analysis). Turbomole[17] and Gaussian 03[18] packages were employed. In DFT

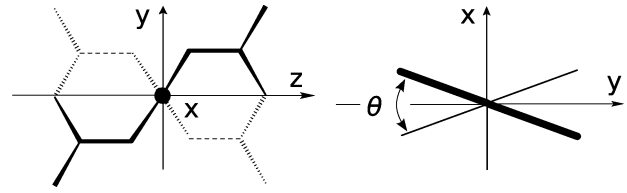


Figure 2. Orientation of $Co(acac)_2$ molecule used in this study: top view (left) and side view (right). The twist angle (θ) is defined as the dihedral angle between the acac rings: zero for the square-planar, 90° for the tetrahedral, and $0 < \theta < 90°$ for a general (twisted) conformation.

Table 1. Definitions of the ANO Basis Sets Used in This Study

basis set	definition
ANO1	Co: ANO-RCC 7s6p5d2f1g, ligands: ANO-S [O: 4s3p1d, C: 3s3p1d, H: 2s]
ANO2	ANO-RCC [Co: 7s6p5d3f2g1h, O: 4s3p2d1f, C: 4s3p1d, H: 3s1p]
ANO3	ANO-RCC [Co: 8s7p6d4f3g2h, O: 5s4p3d2f1g, C: 4s3p2d1f, H: 3s1p]

calculations, two types of basis set were tested: the smaller one (basis A) consists of ecp-10-mdf 6s5p3d1f (with pseudopotential)[19] on Co and 6-31G(d)[20,21] on ligands; the larger one (basis B) consists of QZVPP[22] on Co and TZVPP[23] on ligands. Various density functionals were applied, including (1) generalized gradient approximations (GGA): BP86,[24,25] PBE,[26] OLYP,[27,28] (2) meta-GGA: TPSS,[29] (3) hybrid GGA: PBE0,[30] B3LYP,[31] and (4) hybrid meta-GGA: TPSSH.[32] The influence of solvent on relative energies was estimated using conductor-like screening model (COSMO)[33] as implemented in Turbomole. The dielectric constant (ϵ) of 5.0 was chosen to mimic nonpolar solvents typical to CRP reactions. Apart from that, the infinite ϵ was tested as an upper estimation of nonspecific solvent effects.

CASSCF/CASPT2 energy calculations were performed with the Molcas 6.4 package[34,35] for PBE0 (basis A) structures. The zero-order Hamiltonian for CASPT2 was the default one in Molcas (IPEA shift of 0.25 au[36]) with extra imaginary level shift (0.1 au). Core orbitals were frozen in CASPT2 except 3s and 3p of Co, which were correlated. Three atomic natural orbitals (ANO) basis sets[37,38] of increasing quality, denoted ANO 1−3, were used in the calculations (Table 1). Note that similar basis sets have been successfully employed in CASPT2 calculations for iron complexes.[15,16] Scalar relativistic effects were included in the calculations via second-order Douglas−Kroll−Hess transformation.[39]

The active space was composed of bonding and antibonding orbitals directly participating in the metal−ligand bond, the remaining Co 3d orbitals, and correlating Co 4d (double shell) orbitals. This choice covers the important nondynamical correlation effects in both the cobalt−ligand bond and the cobalt ion itself.[14] It gives in total nine electrons in 11 orbitals (9 in 11) for the planar geometry and 11 electrons in 12 orbitals (11 in 12) for the tetrahedral geometry. The difference is situated in a different number of bonding−antibonding orbital pairs describing the metal−ligand bond: only one such pair exists for the square-planar conformation but

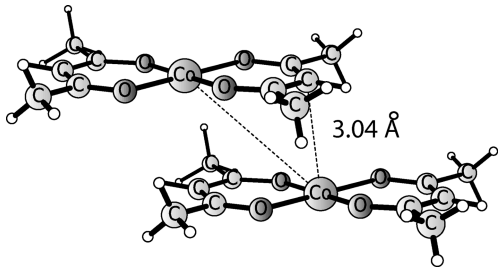


Figure 3. Dimer model, $[Co(acac)_2]_2$.

two (symmetry equivalent) can be found for the tetrahedral form (see contour plots in the Supporting Information). In other words: for the tetrahedral form both Co $3d_{xz}$ and $3d_{yz}$ orbitals efficiently mix with the ligand orbitals, but for the square-planar form only $3d_{yz}$ does so, while the $3d_{xz}$ orbital interacts with the ligand orbitals very weakly, yielding the antibonding combination being almost pure metal orbital and the bonding combination being almost pure ligand orbital (Supporting Information). The latter ligand orbital (with a trace of $3d_{xz}$) was added to the (9 in 11) active space, to obtain the (11 in 12) active space, which is fully compatible with the (11 in 12) active space for the tetrahedral conformation. Nevertheless, it follows that both (11 in 12) and (9 in 11) active spaces lead to almost the same energies for the square-planar conformation (the difference being 0.1−0.8 kcal/mol, below a chemical accuracy). Thus, we conclude that the orbital added in (11 in 12) is indeed not important for the present problem. For these reasons, we recommend using the (9 in 11) active space for the square-planar form, as more intuitive and chemically motivated.

To probe intermolecular interactions in the crystal structure from ref 5, a dimer model was used (Figure 3). The coordinates were taken from the crystal structure and optimized at the PBE0 (basis A) level, constraining the C_{2h} symmetry. The Co−Co′ and Co−C′ distances (between the atoms of the neighboring molecules), which are the structural parameters most important for the stacking interaction, were frozen (in Figure 3 the frozen distances are indicated as dashed lines). The stabilization energies, with single point correction for basis set superposition error (BSSE),[40] are given with respect to isolated monomers in their optimal PBE0/A geometries. Unfortunately, the dimer model is already too large for CASPT2 calculations with the analogous active space as for the monomer. Therefore, the stabilization energies were obtained from Moeller−Plesset second-order perturbation theory (MP2) and approximated second-order coupled clusters model (CC2),[41] both within the RI approximation, as well as from the selected DFT functionals. These calculations employed basis set B. All calculations for the dimer model were performed with the Turbomole package.

For further confirmation of the results from the dimer model, we also carried out analogous RI-MP2/B calculations for the trimer, which is a more realistic model of the crystal structure. For simplicity, the BSSE correction of −2.3 kcal/mol/single molecule was not computed for the trimer, but taken from the dimer calculations (this detail should not be crucial).

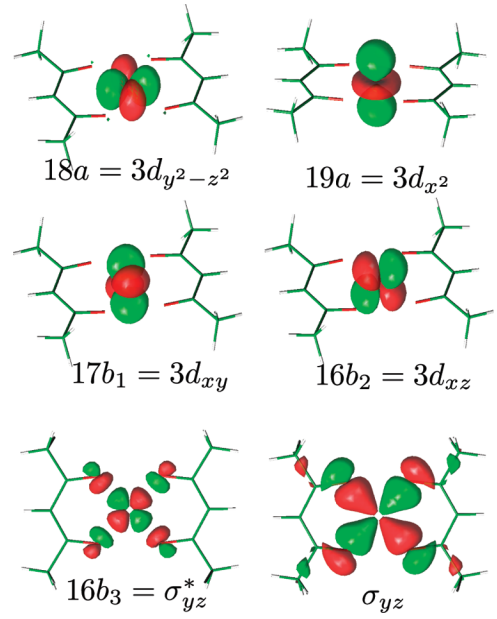


Figure 4. Selected active orbitals (9 in 11) for the square-planar conformation (0.04 $au^{-3/2}$ isovalue).

3. Results and Discussion

3.1. Electronic Configurations and Structures. In the coordination environments of both square-planar and tetrahedral conformations, three of five Co 3d orbitals remain almost pure metal orbital (i.e., do not mix significantly with any ligand orbitals): $3d_{z^2}$ and $3d_{x^2-y^2}$ (18,19a), and $3d_{xy}$ (17b_1). The situation of the remaining two Co 3d orbitals depends on the conformation. For the tetrahedral geometry, both $3d_{xz}$ and $3d_{yz}$ orbitals are destabilized by significant interactions with the corresponding ligand orbitals, yielding $16b_2 = \sigma_{xz}$* and $16b_3 = \sigma_{yz}$* antibonding orbitals (and their bonding counterparts: σ_{xz}, σ_{yz}). For the square-planar geometry, only $3d_{yz}$ is involved in the similar interaction with a single ligand orbital, yielding $16b_3 = \sigma_{yz}$* (and its bonding counterpart: σ_{yz}), while $3d_{xz}$ remains almost pure metal orbital. The contour plots of the discussed orbitals are given in Figures 4 and 5 (the figures show the CASSCF natural orbitals, but DFT orbitals are qualitatively similar).

Now it is useful to take a look at the low-lying electronic states considered in this study (candidates for the ground state, or at least the lowest state for each conformation) and analyze their dominant configurations. For the tetrahedral structure, the lowest quartet state is (regardless the method) $^4A = (18a)^2(19a)^2(17b_1)^1(16b_2)^1(16b_3)^1$ (having unpaired electrons in $3d_{xy}$, σ_{xz}*, and σ_{yz}*). For the square-planar conformation, either $^2A = (18a)^2(19a)^1(17b_1)^2(16b_2)^2(16b_3)^0$ (having an unpaired electron in $3d_{x^2}$) or $^2B_2 = (18a)^2(19a)^2$-$(17b_1)^2(16b_2)^1(16b_3)^0$ (having an unpaired electron in $3d_{xz}$) is the lowest doublet state (depending on the method); therefore, both ^2A and $^2B_2$ will be considered below. As also found in the other study,[6] the tetrahedral geometry does not yield a stable minimum in the doublet state, the geometry always converging to the planar minimum. In other words, the tetrahedral doublets are unstable with respect to twisting of the acac rings (the θ angle, Figure 2), converting one conformation into another. The lowest quartet state for the square-planar conformation is (for all methods) $^4B_1$ =

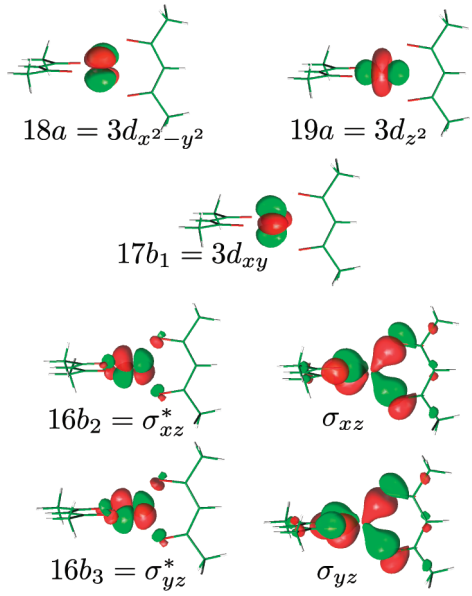


Figure 5. Selected active orbitals (11 in 12) for the tetrahedral conformation (0.04 au$^{-3/2}$ isovalue).

Table 2. Cobalt−Oxygen Distance (in Angstroms)

DFT, $S = {}^1/_2$	1.85−1.90[a]
DFT, $S = {}^3/_2$	1.94−2.00[a]
expt, tetrahedral (tetrameric)[4]	1.93−2.13[b]
expt, square-planar[5]	1.92

[a] Depending on the functional. [b] Nonequivalent bond lengths found in the crystal.

$(18a)^1(19a)^2(17b_1)^2(16b_2)^1(16b_3)^1$, having unpaired electrons in $3d_{y^2-z^2}$, $3d_{xz}$, and σ_{yz}*. This state is also unstable with respect to θ-twisting mode, in agreement with ref 6. However, it was found that another quartet state (of only slightly higher energy) gives a stable minimum for the planar conformation: $^4B_3 = (18a)^1(19a)^1(17b_1)^2(16b_2)^2(16b_3)^1$, having unpaired electrons in $3d_{y^2-z^2}$, $3d_{x^2}$, and σ_{yz}*. Thus, both $^4B_1$ and $^4B_3$ square-planar quartets will be considered.

All tested DFT methods yield very similar geometries for the respective electronic states. (It is generally recognized that different density functionals usually predict very similar structures,[42,43] which is certainly not the case for spin-state energetics.) The representative PBE0/A structures (i.e., those used in CASSCF/CASPT2) can be found in the Supporting Information. Herein we discuss only the key structural parameter, which is the cobalt−oxygen distance (Table 2, DFT values from different functionals are given as ranges for simplicity). The cobalt−oxygen distance, depending mostly on the spin state, is significantly longer in the high-spin than in the low-spin states. This trend is understandable since antibonding metal−ligand orbitals (σ_{yz}*, σ_{xz}*) are occupied in the quartets, but not in the doublets. In a complementary view, the ionic radius of a cobalt cation increases with the increase of spin multiplicity, so the 3d unpaired electrons can better avoid each other in space. The computed Co−O distance for the quartets is close to the bond length characteristic for the oligomeric tetrahedral structure. The experimental value for the other crystallographic structure (with square-planar molecules) is shorter, which would suggest the doublet state. However, this distance falls between the calculated ones for the doublet and the quartet electronic states. For the present case, it is therefore impossible to assign the spin state solely on the basis of the cobalt−oxygen bond length. The analysis of relative energies in the next section attempts to fill this gap.

Table 3. Energies (kcal/mol) of the Square-Planar Doublets and Quartets with Respect to the Tetrahedral ^{4}A

	$S = {}^1/_2$		$S = {}^3/_2$	
	^{2}A	^{2}B$_2$	^{4}B$_1$	^{4}B$_3$
	DFT/Basis A			
PBE0	17.4	20.2	8.6	11.6
B3LYP	12.1	14.4	8.6	12.1
BP86	3.5	1.1	9.3	12.7
PBE	3.8	0.8	8.9	12.5
OLYP	8.5	7.2	8.8	11.4
	DFT/Basis B			
PBE0	13.5	15.7	9.8	12.0
B3LYP	9.8	11.0	9.4	12.0
BP86	−0.5	−3.7	10.7	13.7
PBE	−0.5	−3.9	10.3	13.4
OLYP	5.5	4.0	10.0	12.6
TPSS	0.9	−0.9	11.7	14.7
TPSSH	6.9	7.5	11.1	14.2
	CASPT2/ANO 1−3			
ANO 1	18.2	15.0	9.9	11.2
ANO 2	16.6	13.2	9.0	10.3
ANO 3	16.7	13.2	9.2	10.4

3.2. Relative Energies: CASPT2 vs DFT. DFT and CASPT2 energies of the square-planar doublet (^{2}A, ^{2}B$_2$) and quartet ($^4B_1$, $^4B_3$) states relative to the tetrahedral quartet (^{4}A) state are given in Table 3.

First of all, rather strong dependence of DFT results on the basis set can be observed in agreement with some previous studies.[16,44] Unless otherwise stated, we refer to the results from the larger basis. It can be seen that hybrid density functionals (B3LYP, PBE0) predict a tetrahedral quartet (^{4}A) ground state, being separated from the square-planar doublets and quartets by a clear gap of at least 10 kcal/mol. On the other hand, classical GGA functionals (BP86, PBE) yield a completely different picture with the lowest doublet and quartet close in energy (separated by ∼4 kcal/mol or less). In the larger basis set, the planar doublet eventually becomes the ground state. Meta-GGA functional (TPSS) behaves very similarly to GGA in this respect. Interestingly, somewhere between these extremes there are two other functionals: OLYP (a GGA based on the OPTX exchange[27]) and TPSSH (hybrid meta-GGA, containing only 10% of Hartree−Fock exchange). It is clear that the doublet-quartet splitting in Co(acac)$_2$ strongly depends on the applied functional, particularly on the admixture of Hartree−Fock exchange in the exchange potential. A qualitatively similar trend was observed in several DFT studies (see refs 15 and 45 and references therein). It is worth noting here that the (adiabatic) DFT doublet−quartet splitting for Co(acac)$_2$ ranges from −3.9 (PBE) to 13.5 (PBE0). This uncertainty is a warning that DFT calculations cannot be safely used to determine the ground-state multiplicity of Co(acac)$_2$.

The CASPT2 calculations predict the high-spin ^{4}A (in the tetrahedral conformation) to be the ground state. The lowest doublet (^{2}B$_2$) lies about 13 kcal/mol above; the lowest square-

Table 4. Doublet−Quartet Splitting in Free Co(II) Ion

		ΔE (^{2}G− ^{4}F) (cm^{-1})	error (kcal/mol)
6s5p4d2f1g		16651	0.4
7s6p5d2f1g	(ANO 1)	16488	−0.1
7s6p5d3f2g1h	(ANO 2)	16258	−0.7
8s7p6d4f3g2h	(ANO 3)	16140	−1.1
exptl[a]	(NIST[47])	16510	0

[a] Average energy of spin−orbit terms.

planar quartet ($^4B_1$) lies about 9 kcal/mol. These results are stable as the basis set is increased: going from ANO 2 to ANO 3 has only a minor impact on the relative energies, and therefore the results seem to be converged with respect to basis set. To further analyze the basis set dependence, we investigated doublet−quartet (^{2}G−^{4}F) splitting in the free cobalt(II) ion, a property that is known experimentally with high accuracy. As argued elsewhere, the basis set error on the $Co(acac)_2$ doublet−quartet splitting can be partially traced back to the analogous error on the atomic excitation energy.[15,46] The experimental data was obtained from the National Institute of Science and Technology (NIST) Atomic Spectra Database[47] by averaging the energies of spin−orbit terms according to the procedure described elsewhere.[46] The theoretical CASPT2 results were computed with 3d4d active space for various contractions of the ANO-RCC basis set, corresponding to ANO 1−3 (Table 4). It follows that for free Co(II) ion CASPT2 is accurate up to ∼1 kcal/mol for all the tested basis sets, even the smallest one. (It might be interesting that starting from ANO 1 the error slightly increases as basis set is increased; obviously, this behavior is rooted in an approximate character of the CASSCF/CASPT2 method itself.) The error falls within chemical accuracy and is fully acceptable for all cases. Interestingly, in this atomic calculations the CASPT2 method tends to slightly favor the low-spin state; it does not reveal any tendency to overstabilize the high-spin state. This observation supports the previous conclusion about the quartet ground state.

CASPT2 state ordering can be satisfactorily reproduced only by hybrid functionals (B3LYP, PBE0). Nonhybrid functionals (maybe except OLYP) artificially favor the low-spin states and thus yield different (and probably wrong) results. On the other hand, all tested functionals accurately reproduce the energy difference between the planar and the tetrahedral conformations for the high-spin states. While DFT apparently has problems with spin-state energetics, these problems do not affect the energy of a simple conformational transition not accompanied by a spin flipping.

Thus far, only electronic energy differences were discussed. Two other factors that might discriminate between alternative electronic states are vibrational energies and solvent effects. Their impact on relative energies is described in the Supporting Information. Here, we summarize that the addition of zero-point vibrational energies further increases (by ∼1 kcal/mol) the relative energies of the square-planar doublets, while it has much smaller influence on the relative energies of the square-planar quartets (with respect to the tetrahedral quartet). Thus, considering vibrational energies by no means alters our previous conclusion about the ground state; conversely: the high-spin ground state is stabilized even more. On the other hand, the polarizable continuum stabilizes the square-planar conformation by 1−2 kcal/mol with respect to the tetrahedral one. This influence is definitely too small to change the ground-state identity. Therefore, we do expect the same ground state in vacuum as in an apolar, noncoordinating solvent.

Table 5. Stabilization Energy (kcal/mol) per Single $Co(acac)_2$ Molecule Estimated from the Dimer Model

	doublet	quartet
RI-MP2	9.5	9.5
RI-CC2	11.1	10.2
HF	1.2	2.2
PBE0	2.9	3.0
PBE	2.4	3.7
BP86	0.8	−0.4

3.3. Stacking Energy in the Crystal Structure. According to the above CASPT2 calculations, the high-spin state in tetrahedral conformation should be preferred by at least 10 kcal/mol. Nevertheless, the recently published crystal structure consisting of square-planar molecules[5] appears to contradict this result, which calls for explanation. To this end, we focused on prospective intermolecular interactions favoring the planar conformation. Indeed, it was pointed out that in the crystal structure the molecules are loosely stacked on top of each other,[6] and thus the resulting stacking interactions (weak "sandwich" coordination of cobalt by π electrons of the neighboring acac ring and/or dispersion) might stabilize the planar form, explaining its presence in the crystal. However, if it were the correct explanation, the stabilization should be stronger than 10 kcal/mol/single molecule (the amount of energy by which the tetrahedral conformation is preferred in vacuum), which seems quite large. Please note that the distance from Co to the nearest neighbor acac ring is ∼3 Å (cf. Figure 3 and ref 5). Therefore, we attempted to assess the strength of the stacking interaction at a quantitative level.

To this end, we used the dimer model representing the considered crystal structure (details in section 2). The interaction energies were computed with respect to the doublet (^{2}A) and quartet ($^4B_1$) states of monomers. Thereby, two spin states were considered for the dimer: the triplet (i.e., the dimer of doublets) and the septet (i.e., the dimer of quartets). As expected, the stabilization energies were very similar for both spin states. The computed stabilization energies per single molecule are given in Table 5 (note that a positive sign of stabilization energy indicates stabilization, and negative indicates destabilization).

Single reference ab initio methods (MP2, CC2) yield a very similar and quite large dimer stabilization energy, on the order of 20 kcal/mol, i.e., 10 kcal/mol for a monomer. It is larger than one would expect for "typical" intermolecular interactions. The known cobalt complexes where some ligands are loosely bound to Co via delocalized π electrons[10] should be recalled here. Likewise, the delocalized π density of the acac ligand can coordinate the cobalt ion from the neighboring molecule. Let us note, however, that the intermolecular distance is large (∼3 Å). Therefore, we do not attempt to firmly quantify the role of chemical interaction or just dispersion forces in the net stabilization energy. A

more systematic treatment, such as symmetry-adapted perturbation theory,[48] would be desirable for this purpose, but such a study is beyond the scope of the present investigation. Let us note, however, that HF and the tested DFT methods predict only very small stabilization and — as these methods do not account properly for dispersion effects, while MP2 and CC2 do — this could indicate a dispersion as an important ingredient of stabilization. We do not rule out, however, the possibility of a weak chemical bond with π electrons of the acac ring. Given the poor performance of the tested DFT methods for the stabilization energy (Table 5), we suspect that recently developed hybrid meta-GGA functionals, demonstrated to better reproduce stacking energies in organic systems,[49] might be somewhat better suited for this task.

By definition, the stacking interaction discussed above affects only the planar conformation. Certainly, some intermolecular interaction would also affect the tetrahedral conformation, but this interaction should presumably be much smaller. Roughly speaking, it is the planarity of the $Co(acac)_2$ molecule that warrants its strong, favorable interaction with the other (also planar) molecule. We believe therefore that the rigorous assessment of weak intermolecular interactions affecting the tetrahedral structure (a very demanding project, including the search for the most favorable crystal structure with tetrahedral cobalt) is not obligatory here. Thus, the stabilization gained by the square-planar conformation due to stacking interaction could indeed compensate for its intrinsically higher energy, which could explain (at a semiquantitative level) the apparent discrepancy between theoretical geometry of a monomer and the crystallographic structure from ref 5. Nevertheless, higher-order ab initio calculations of stabilization energies, for instance with CCSD or CCSD(T), would be desirable for further confirmation. These calculations are beyond our computational power, though.

It is worth noting here that the actual stabilization energies (in the frame of MP2 and CC2 methods) could be even higher than those computed here. First, this could be because of the applied correction for basis set superposition error, which is often claimed to make binding energies too low. Second, the dimer model is likely to give a too-small stabilization energy compared to that of the crystal lattice (assuming full additivity of the nearest neighbor interactions, the dimer model would give only a half of the actual stabilization energy). However, the main drawback of the dimer model is that the $Co(acac)_2$ unit has only one neighbor (from one side), while in solid it has two neighbors (from both sides). In this respect, the trimer model would be a much better approximation. We have therefore performed RI-MP2 calculations for the trimer model, which gave a stabilization energy of 13.5 kcal/mol per single molecule, even larger than that for the dimer model, which supports our conclusions.

Let us now hypothesize on the spin state of $Co(acac)_2$ present in the lattice reported in ref 5. According to the above CASPT2 results (cf. Table 3), the quartet ($^4B_1$) is the lowest spin state for the square-planar conformation, the lowest doublet state lying ~4 kcal/mol above. We also saw that the stacking interaction does not discriminate between doublet and quartet states of monomers (cf. Table 5). Therefore, most probably the high-spin state is "frozen" inside the crystal. This prediction seems also supported by a low-temperature, solid-state EPR spectrum registered by Sojka et al.[7] The suspected presence of a square-planar quartet state in the crystal cannot be diminished by its conformational instability reported in section 3.1. First, the other quartet state $^4B_3$, being perfectly stable at square-planar geometry, lies only ~1 kcal/mol above $^4B_1$. Such closely lying states could be, in fact, strongly coupled via a spin−orbit term (not accounted for here), losing their exact orbital symmetry. Second, the instability for $^4B_1$ is actually very weak. To check its importance, we computed potential energy curves (in a function of θ) for $^4B_1$ and the other relevant states (Figure 6). Evidently, the energy curve for $^4B_1$ is almost flat, indicating only a negligible instability, which could be easily hindered in a solid, particularly here, in the presence of strong π-stacking interactions.

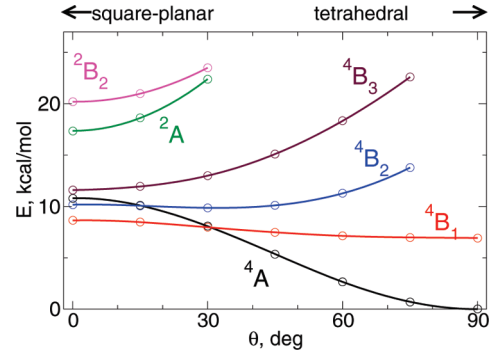


Figure 6. Electronic energy curves (PBE0, basis A) as a function of the twist angle θ. The $^4B_1$ and $^4B_2$ states are (negligibly) unstable at $\theta = 0$, having instead shallow minima at $\theta = 90°$ and $\theta \approx 30°$, respectively.

4. Conclusions

The ground state of bis(acetylacetonate) cobalt(II) was identified as high-spin and tetrahedral by means of ab initio CASPT2 calculations with a high credibility level. In variance, DFT results for $Co(acac)_2$ substantially depend on the exchange functional. The comparison with CASPT2 supports the results from the hybrid functionals with 20−25% of HF exchange (B3LYP, PBE0). Nonhybrid ones (BP86, PBE, OLYP, TPSS) overstabilize the low-spin state and are in disagreement with CASPT2 by 9−10 kcal/mol or more, in some cases (BP86, PBE) predicting even the wrong spin state and the wrong conformation. Also, the OLYP functional, supported by several successful comparisons with experiment or ab initio calculations,[16,50,51] is herein quite far from CASPT2 (though not as far as classical GGA: BP86 and PBE).

The conclusions about DFT performance, drawn here by comparison with CASPT2, have a benchmark value and can be used to choose the optimal functional for similar compounds (e.g., with substituted acac ligands). The wise choice of density functional is particularly important for modeling CRP and other reactions involving $Co(acac)_2$ or derivatives. In the studies on reaction mechanisms, the big size and low symmetry of the models and other technical

difficulties usually make ab initio calculations prohibitively expensive, thus nolens volens one must stick to DFT. Thus, given the CASPT2 results, we recommend the hybrid functionals for the present and similar cases. Note, however, that this recommendation cannot be universal, being necessarily limited to ligand fields similar to those studied here. For instance, the hybrid functionals might not be optimal to describe coordination of an extra ligand to $Co(acac)_2$; this issue would require further studies. As already evident from many examples,[11,15,16] DFT performance for spin-state energetics shows to be variable and capricious.

We also focused on the crystal structure of monomeric bis(acetylacetonate) cobalt(II) from ref 5 (containing square-planar molecules) in view of its putative contradiction with our calculations. Therefore, we estimated the energy of the intermolecular interactions in the crystal structure at MP2 and CC2 levels using dimer and trimer models. We found that the favorable interaction "felt" by a square-planar monomer might be strong enough ($\gtrsim$10 kcal/mol) to compensate for its intrinsically higher energy. This result supports the original hypothesis given in ref 6 at quantitative level and could explain the apparent discrepancy between theory (CASPT2, hybrid DFT) and the crystal structure.[5] According to the relative spin state energetics presented here, most probably the high-spin state is "frozen" in the crystal structure.

It should be stressed here that BP86 or PBE results, predicting a planar conformation for an isolated $Co(acac)_2$ molecule, could be naively treated as "confirmed" in view the crystal structure from ref 5. In fact, the optimal conformation of a single molecule is nonplanar, but the crystal structure is dominated by strong π-stacking interactions. Nota bene, BP86 and PBE are not able to properly describe these interactions. This is a methodological warning that the possibility of strong intermolecular interactions must be accounted for when judging the accuracy of quantum chemical methods on the basis of crystal structures.

Acknowledgment. We are thankful to Dr. Artur Michalak (Jagiellonian University) for focusing our attention on methodological problems with cobalt(II) bis(acetylacetonate). We also acknowledge Prof. Zbigniew Sojka and Prof. Piotr Petelenz (Jagiellonian University) for interesting discussions and Prof. Kristine Pierloot (University of Leuven) for valuable private communication. Special thanks to her Ph.D. student, Steven Vancoillie, for constructive remarks about the manuscript. This work was partially supported by computational grants from the Academic Computer Center CYFRONET AGH (Grant Nos.: MNiSW/SGI3700/UJ/133/2006, MNiSW/SGI4700/UJ/085/2007, and MNiSW/IBM_BC_HS21/UJ/085/2007). Molecular structures in Figures 1 and 3 were prepared with the XYZViewer program, courtesy of its author, Sven de Marothy (Stockholm University, 2003−2004).

Supporting Information Available: The PBE0/A structures, the contour plots of the orbitals discussed in section 2, and the estimation of solvent effects and relative zero-point vibrational energies. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kaneyoshi, H.; Matyjaszewski, K. *Macromolecules* **2005**, *38*, 8163–8169.

(2) Cotton, F. A.; Holm, R. H. *J. Am. Chem. Soc.* **1960**, *82*, 2979–2983.

(3) Cotton, F. A.; Soderberg, R. H. *Inorg. Chem.* **1964**, *3*, 1–5.

(4) Cotton, F. A.; Elder, R. C. *Inorg. Chem.* **1965**, *4*, 1145–1151.

(5) Burgess, J.; Fawcett, J.; Russell, D. H.; Gilani, S. R. *Acta Crystallogr.* **2000**, *C56*, 649–650.

(6) Maria, S.; Kaneyoushi, H.; Matyjaszewski, K.; Poli, R. *Chem.—Eur. J.* **2007**, *13*, 2480–2492.

(7) Srebro, M.; Pietrzyk, P.; Michalak, A.; Sojka, Z. To be submitted for publication.

(8) Nindakova, L.; Shainyan, B.; Saraev, V.; Chipanina, N.; Umanetz, V. *J. Mol. Catal. A: Chem.* **2005**, *235*, 161–172.

(9) Debuigne, A.; Champouret, Y.; Jérôme, R.; Poli, R.; Detrembleur, C. *Chem.—Eur. J.* **2008**, *14*, 4046–4059.

(10) Debuigne, A.; Michaux, C.; Jérôme, C.; Jérôme, R.; Poli, R.; Detrembleur, C. *Chem.—Eur. J.* **2008**, *14*, 7623–7637.

(11) Harvey, J. N. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2006**, *102*, 203–226.

(12) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157–173.

(13) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1991**, *96*, 1218–1226.

(14) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094.

(15) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2006**, *125*, 124303.

(16) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2008**, *128*, 034104.

(17) Ahlrichs, R.; Horn, H.; Schaefer, A.; Treutler, O.; Haeser, M.; Baer, M.; Boecker, S.; Deglmann, P.; Furche, F. Turbomole, version 5.9; Universität Karlsruhe: Karlsruhe, Germany.

(18) 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 C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(19) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866–872.

(20) Hehre, W.; Ditchfield, R.; Pople, J. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

(21) Francl, M.; Petro, W.; Hehre, W.; Binkley, J.; Gordon, M.; DeFrees, D.; Pople, J. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(22) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753–12762.

(23) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(24) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(25) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(26) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(27) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.

(28) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(29) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401–146404.

(30) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(32) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

(33) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(34) Andersson, K.; Barysz, M.; Bernhardsson, A.; Blomberg, M.; Cooper, D.; Fülscher, M.; de Graaf, C.; Hess, B.; Karlstrom, G.; Lindh, R.; Malmqvist, P.-Å.; Nakajima, T.; Neogrady, P.; Olsen, J.; Roos, B.; Schimmelpfennig, B.; Schütz, M.; Seijo, L.; Serrano-Andrés, P. E. M.; Siegbahn, J. S.; Thorsteinsson, T.; Veryazov, V.; Widmark, P.-O. MOLCAS, version 6.4, Lund University: Lund, Sweden, 2006.

(35) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(36) Ghigo, G.; Roos, B.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149 .

(37) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. *Theor. Chim. Acta* **1995**, *90*, 87–114.

(38) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.

(39) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 10945–10956.

(40) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(41) Christiansen, O.; Koch, H.; Jorgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

(42) Conradie, J.; Quarless, D.; Hsu, H.-F.; Harrop, T.; Lippard, S.; Koch, S.; Ghosh, A. *J. Am. Chem. Soc.* **2007**, *129*, 10446–10456.

(43) Neese, F. *J. Biol. Inorg. Chem.* **2006**, *11*, 702–711.

(44) Güell, M.; Luis, J. M.; Solà, M.; Swart, M. *J. Phys. Chem. A* **2008**, *112*, 6384–6391.

(45) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48–55.

(46) Fouqueau, A.; Mer, S.; Casida, M. E. *J. Chem. Phys.* **2004**, *120*, 9473–9486.

(47) National Institute of Science and Technology (NIST) Atomic Spectra Database. http://www.physics.nist.gov/cgi-bin/AtData/main_asd (Accessed December 2008).

(48) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

(49) Zhao, Y.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2701–2705.

(50) Conradie, J.; Ghosh, A. *J. Phys. Chem. B* **2007**, *111*, 12621–12624.

(51) Aquilante, F.; Malmqvist, P.-Ã.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694–702.

CT800571Y

JCTC Journal of Chemical Theory and Computation

Molecular Orbital-Averaged Fukui Function for the Reactivity Description of Alkaline Earth Metal Oxide Clusters

Nick Sablon,* Frank De Proft, and Paul Geerlings

Vrije Universiteit Brussel, Eenheid Algemene Chemie, Pleinlaan 2, B-1050 Brussel, Belgium

Received January 16, 2009

Abstract: This paper concerns the accurate description of the surface oxygen reactivity for the alkaline earth metal oxides using local DFT-based reactivity indices. The cumbersome periodic boundary conditions calculations, typically required to probe the reactivity of such systems, are avoided by the construction of a reliable cluster model. The standard Fukui function concept of conceptual DFT is generalized to include the contribution from not only the HOMO and the LUMO but also from other chemically relevant orbitals. Results prove that this approach is a valuable and straightforward alternative to the reactivity calculation of extended systems.

1. Introduction

In general, chemical reactions are driven by a change in the reagents' number of electrons and a rearrangement of their nuclei. Conceptual density functional theory[1−4] (DFT) exploits this observation to define various reactivity descriptors for the interpretation of chemical reactions. First and higher order (functional) derivatives of the electronic energy with respect to the electron number N or the external potential $v(\mathbf{r})$, which is just the electron−nuclear potential for isolated atoms and molecules, are identified with quantities of chemical relevance. Concepts such as the electronic chemical potential or the electronegativity,[5] the chemical hardness and softness,[6] the Fukui function,[7,8] etc., could sharply be defined in this context and have been applied in many studies to interpret both theoretical and experimental data.[3]

Nowadays the practical computation of these reactivity descriptors usually occurs in a Kohn−Sham (KS) DFT scheme,[9] in which the introduction of the so-called Kohn−Sham molecular orbitals (KS MOs) is key in the efficient application of the variational procedure for the electron density.[10] It turns out that the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) are the primary contributors to the characteristics of the various DFT-based concepts.[11−13] In this perspective, conceptual DFT can to some extent be seen as a generalization of the frontier molecular orbital (FMO) theory taking care of orbital relaxation effects and broadening the range of concepts.

Although it is clear that the HOMO and LUMO play a decisive role in the description of molecules with well-separated MO energy levels, is it also the case when these energy levels are closely spaced and an intermediate situation between the discrete MO levels and the continuous bands occurs? This problem will be addressed in the current paper. We will show that the prominent role of the HOMO and the LUMO should in certain cases be reassigned to and distributed among one or more other orbitals which leads to the introduction of the concept of the MO-averaged Fukui function.

This approach, which can be considered as an approximation to the local density of states-based Fukui function for extended systems, will be tested in a study of the (100) surface reactivity of the alkaline earth metal oxides MgO, CaO, SrO, and BaO. These systems are of particular interest because of their catalytic characteristics and their use in various technological applications. They are for example used in the oxidation process of methane[14] and for the destructive adsorption of chlorinated hydrocarbons.[15] Calculation of the reactivity of the electron rich surface oxygen centers is the main goal of this paper since these atoms are highly reactive toward electophilic attacks[16−18] and could be considered as superbasic sites.[19] The central idea is to use the well-established conceptual DFT framework for mol-

* Author to whom correspondence should be addressed. E-mail: nsablon@vub.ac.be.

10.1021/ct9000312 CCC: $40.75
Published on Web 04/10/2009

ecules in this study in order to avoid the cumbersome infinite system calculations. Even though a simple diatomic gas-phase system could give some reactivity indications, the surface characteristics are inaccessible by this kind of simplified model system. It has, for example, been observed that reactivity trends for bulk and surface atoms of the same systems can be different.[20] This is the reason why we have opted for the use of an embedded cluster approach to construct a rigorous surface model.

The paper is organized as follows. In Section 2 the necessary elements of conceptual DFT are summarized and the MO-averaged Fukui function concept is introduced. Section 3 briefly reviews some of the approaches to construct reliable cluster models, explains the model used, and analyzes the properties of the alkaline earth oxide clusters. Methodological details and the surface reactivity results are discussed in the fourth section. The final conclusions are given in Section 5.

2. Theoretical Background

2.1. The Fukui Function. As this paper primarily concerns the local description of chemical reactivity, global reactivity or stability indicators will not be discussed in detail here. We will, however, give some theoretical information on the Fukui function $f(\mathbf{r})$, which is the concept of interest in a local description of reactions that are driven by the so-called soft−soft interactions. Parr and Yang[7] defined this concept as the second derivative of the electronic energy E with respect to the number of electrons N and the external potential $v(\mathbf{r})$:

$$f(\mathbf{r}) = \frac{\partial^2 E}{\delta v(\mathbf{r}) \partial N} \tag{1}$$

Using the following relations for the electron density $\rho(\mathbf{r})$ and the electronic chemical potential μ,

$$\rho(\mathbf{r}) = \left(\frac{\delta E}{\delta v(\mathbf{r})}\right)_N \tag{2}$$

$$\mu = \left(\frac{\partial E}{\partial N}\right)_{v(\mathbf{r})} \tag{3}$$

two equivalent expressions for the Fukui function can be found:

$$f(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N}\right)_{v(\mathbf{r})} = \left(\frac{\delta \mu}{\delta v(\mathbf{r})}\right)_N \tag{4}$$

The derivatives with respect to N in eqs 1, 3, and 4 imply that one must introduce a noninteger number of electrons. This was done by Perdew et al.[21] in a zero temperature grand canonical ensemble framework, which showed that for exact solutions to the Schrödinger equation the electronic energy, like the electron density evaluated at a certain point in space, as a function of N is made up of a series of straight line segments that intersect at integer electron number.[22,23] These derivatives should consequently be replaced by right- and left-hand side derivatives at the integer electron number. Within an exact theory the chemical potentials and the Fukui functions from above, μ^+ and $f^+(\mathbf{r})$, and from below, μ^- and $f^-(\mathbf{r})$, are thus given by

$$\mu_K^+ = \left(\frac{\partial E}{\partial N}\right)_{v(\mathbf{r})}^+ \Bigg|_{N=K} = E_{K+1} - E_K \tag{5}$$

$$\mu_K^- = \left(\frac{\partial E}{\partial N}\right)_{v(\mathbf{r})}^- \Bigg|_{N=K} = E_K - E_{K-1} \tag{6}$$

$$f_K^+(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N}\right)_{v(\mathbf{r})}^+ \Bigg|_{N=K} = \rho_{K+1}(\mathbf{r}) - \rho_K(\mathbf{r}) \tag{7}$$

$$f_K^-(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N}\right)_{v(\mathbf{r})}^- \Bigg|_{N=K} = \rho_K(\mathbf{r}) - \rho_{K-1}(\mathbf{r}) \tag{8}$$

at an integer number of electrons K. Equation 7 allows the interpretation of $f_K^+(\mathbf{r})$ to describe the propensity of a molecular site to undergo nucleophilic attacks, while eq 8 shows that $f_K^-(\mathbf{r})$ is able to describe the reactivity toward electrophilic attacks.

These finite difference formulas (eqs 5−8) are however not exact anymore when practical DFT calculations with approximate exchange-correlation functionals are carried out.[8] The correct behavior of the Fukui functions $f_K^{\pm}(\mathbf{r})$ within such an approximate theory can thus only be obtained through a direct evaluation of the derivative in eq 1. Based on the fact that DFT calculations are rooted in the KS approach and using Janak's theorem,[24]

$$\left(\frac{\partial E}{\partial n_i}\right) = \varepsilon_i \tag{9}$$

where n_i and ε_i denote the occupation number and the energy of the ith MO, the derivatives of E with respect to N are given by the KS HOMO and LUMO energies,

$$\left(\frac{\partial E}{\partial N}\right)_{v(\mathbf{r})}^+ = \varepsilon_{\mathrm{LUMO}} \tag{10}$$

$$\left(\frac{\partial E}{\partial N}\right)_{v(\mathbf{r})}^- = \varepsilon_{\mathrm{HOMO}} \tag{11}$$

which implies the following expressions for the Fukui functions:

$$f^+(\mathbf{r}) = \left(\frac{\delta \varepsilon_{\mathrm{LUMO}}}{\delta v(\mathbf{r})}\right)_N \tag{12}$$

$$f^-(\mathbf{r}) = \left(\frac{\delta \varepsilon_{\mathrm{HOMO}}}{\delta v(\mathbf{r})}\right)_N \tag{13}$$

(Equations 10 and 11 also hold for an exact KS theory if $\varepsilon_{\mathrm{LUMO}}$ and $\varepsilon_{\mathrm{HOMO}}$ are replaced by $\varepsilon_{\mathrm{LUMO}}^+$ and $\varepsilon_{\mathrm{LUMO}}^-$, which are the frontier orbital energies associated with the exact exchange-correlation potential from above $v_{xc}^+(\mathbf{r})$ and from below $v_{xc}^-(\mathbf{r})$, respectively.[21,22,25] The difference between the two exchange-correlation potentials disappears for local approximations to the exchange-correlation functional (LDA and GGA) so that eqs 10 and 11 can be used.)

So, the Fukui functions can readily be obtained by evaluation of these functional derivatives, for which a methodology is explained in Section 4.

Equations 12 and 13 are clear examples of the importance of the HOMO and the LUMO in conceptual DFT. Other exact relations for the Fukui functions, obtained by Yang et al., are a further illustration:[1,7,11,26]

$$f^{+}(\mathbf{r}) = \sum_{s}\left\{|\psi_{\mathrm{LUMO}}(\mathbf{x})|^2 + \sum_{i=1}^{N}\left(\frac{\partial|\psi_i(\mathbf{x})|^2}{\partial N}\right)^{+}_{\nu(\mathbf{r})}\right\} \quad (14)$$

$$f^{-}(\mathbf{r}) = \sum_{s}\left\{|\psi_{\mathrm{HOMO}}(\mathbf{x})|^2 + \sum_{i=1}^{N}\left(\frac{\partial|\psi_i(\mathbf{x})|^2}{\partial N}\right)^{-}_{\nu(\mathbf{r})}\right\} \quad (15)$$

with $\psi_i(\mathbf{x})$ the KS MOs, depending on a position coordinate $\mathbf{r}$ and a spin coordinate s. The main contribution to the Fukui functions can be ascribed to the probability density of the frontier MOs. Relaxation effects of all the orbitals are, nonetheless, taken into account, which is an improvement to the FMO theory.

2.2. The MO-Averaged Fukui Function. The dominant role the HOMO and the LUMO play in the description of chemical reactivity can be traced back to the underlying assumption that when electron transfer occurs during the chemical reaction the removal or addition of electrons take place in these orbitals. Although this has been shown to be a valid approach in very many cases,[3] for systems presenting a number of closely lying frontier orbitals this might pose serious problems. In the past this has, for example, led to approaches that consider the HOMO-1 orbital in order to probe the reactivity toward electrophiles.[27] The HOMO and LUMO are also generally localized on the most reactive site; even so, one can be interested in a reactivity description of another molecular part, where these orbitals are not necessarily well represented, as it could be more prone to reaction due to steric effects or when a comparison with an analogous site in another molecule is the aim. We recently observed for example that the directing effect of substituents in the electrophilic aromatic substitution reactions of simple monosubstituted benzenes could not always be recovered by the information contained in the HOMO in case of it predominantly being located on the substituent and not on the aromatic ring.[27] For larger systems having a number of closely spaced MO energy levels (which could be seen as a precursor of the band structure in infinite systems), it is irrelevant to pick just one orbital out for the reactivity description while ignoring the others.

These observations suggest that the classical Fukui function concept as introduced above should be extended in certain cases. Other orbitals should be able to enter eqs 12−15 on the same footing as the frontier orbitals. We propose two criteria to rigorously determine the appropriate orbitals and their contribution to the reactivity description, which will lead to the definition of the MO-averaged Fukui function ($\tilde{f}^{+}(\mathbf{r})$ or $\tilde{f}^{-}(\mathbf{r})$) concept.

The first one is an energetic criterion. It is clear that only orbitals which have the energetic potential to exert a substantial chemical influence (which have an energy comparable to the HOMO or LUMO energy) should be considered. (It is interesting to note that, in the case where the orbital energy spacing is very small, energetic responses induced by perturbations in the external potential are likely to exceed this spacing so that the Fukui functions are only strictly defined within a degenerate perturbation theoretical framework. The averaging proposed in eqs 16 and 17 is, however, a more straightforward approach which includes the responses of the necessary orbitals for a reliable reactivity description. This observation was brought up by an anonymous referee, who the authors wish to acknowledge.) A close inspection of the MO energy levels is thus required in order to select the relevant orbitals. A spatial criterion, on the other hand, demanding that the included orbitals are particularly situated on the reactive site of interest is indispensable when an intermolecular reactivity scale of a specific molecular site is desired. This can be assured by weighting the various orbital contributions with a factor that is proportional to the orbitals' probability density at this site. The MO-averaged Fukui functions summarize all these requirements and are defined as

$$\tilde{f}^{+}(\mathbf{r}) = \sum_{i} w_i f_i^{+}(\mathbf{r}) \quad (16)$$

$$\tilde{f}^{-}(\mathbf{r}) = \sum_{i} w_i f_i^{-}(\mathbf{r}) \quad (17)$$

where w_i are the aforementioned weighting factors and where the index i in eq 16 runs over all the unoccupied MOs satisfying the energetic criterion, while it considers the occupied ones in eq 17. It should be stressed that the inclusion of the weighting factors is not to be done if one wants to discern the various reactive regions within a molecule and compare their relative importance. $f_i^{+}(\mathbf{r})$ and $f_i^{-}(\mathbf{r})$ denote alternative Fukui functions which are analogous to the ones in eqs 12 to 15, apart from the fact that the LUMO or the HOMO are exchanged for the ith unoccupied or occupied MO respectively, and reflect the reactive behavior of this orbital with respect to a nucleophilic or electrophilic attack. The introduction of these alternative Fukui functions is in total agreement with Janak's theorem if one assumes that the electron addition or withdrawal indeed occurs in these orbitals.

An explicit expression for the calculation of the weighting factors w_i will be presented in Section 4, where the dramatic improvement of the reactivity description by application of eq 17 will be extensively discussed as well. But before this can be done, we will elaborate on the kind of systems to be analyzed and more specifically on the choice of the cluster model for the alkaline earth oxides and on the electronic structure of the reactive site of interest.

3. Cluster Models and Analysis of the Alkaline Earth Oxide Clusters

The modeling of infinite systems as solids and solid surfaces by finite clusters has been discussed in a vast amount of papers. Often metal oxides are concerned because of the interest in their reactivity properties, which include adsorption characteristics, crystal defects, electronic structure, and catalytic activity. Many approaches have been proposed to construct appropriate clusters that accurately simulate the genuine solid state and to improve their performance. The

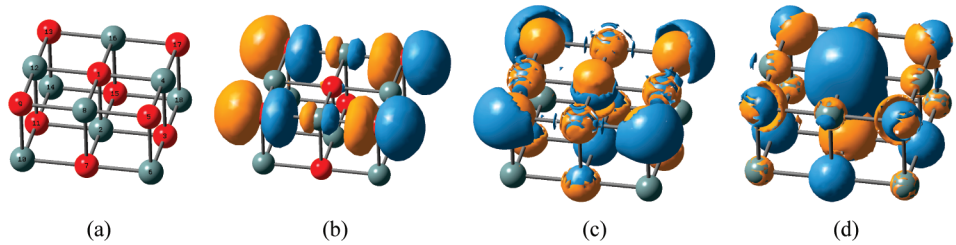


Figure 1. Sr_9O_9 cluster. (a) Representation of the cut-out cluster with atom 1 at the center as the surface oxygen of interest. (b) Isosurface of the HOMO with an isovalue of 0.02 au. (c) Isosurface of the Fukui function from below $f^-(\mathbf{r})$, calculated with eq 13, with an isovalue of 0.002 au. (d) Isosurface of the MO-averaged Fukui function from below $\tilde{f}^-(\mathbf{r})$, calculated with eq 17, with an isovalue of 0.002 au.

first paragraph of this section briefly reviews the most common ones in order to justify our model chosen. A discussion of the whole panoply of available methods lies beyond the scope of this paper. The second paragraph applies our model of choice to the calculation of MgO, CaO, SrO, and BaO and analyses their cluster properties.

3.1. Cluster Models. In the case of metal oxides, a subdivision can be made between the bare, embedded, and saturated cluster models.[28] A bare cluster only consists of a piece of the bulk structure, and its relevance is often strongly dependent on a fortuitously good choice of geometry and size.[29] It is clear that an improved description results from incorporating the effect of the rest of the bulk, which is not explicitly taken into account in the quantum chemical calculations. For ionic solids, as the alkaline earth oxides, an embedding procedure which mimics the potential created by the rest of the bulk by surrounding the cluster with an array of point charges is the most common and effective approach.[30,31]

As the embedding potential is a correction to the bare cluster for the neglected bulk effects (i.e., the electrostatic or Madelung potential and boundary interactions), the cut-out should rigorously be done so as to represent the bulk situation as accurately as possible. Lü et al.[31,32] defined three practical rules that ensure an easy and good choice. These are the stoichiometry, neutrality, and coordination principles. The use of a stoichiometric cluster seems to improve results because spurious energy levels coming from excess atoms are avoided.[33] Moreover, the neutrality condition will then automatically be fulfilled. In order to be in line with the coordination principle, a cluster should contain the strongest bonds within it.

Various embedding schemes have been developed. As indicated above, methods based on cluster enclosure in an array of point charges are prevalent, though other possibilities such as the hybrid quantum mechanical/molecular mechanical (QM/MM) treatment exist.[34] The easiest approach is to place point charges at the ideal lattice positions and assign them values equaling the ionic charges of the crystal. This model can be refined by replacing the positive charges closest to the cluster by pseudopotentials.[35] An overpolarization of the neighboring cluster anions is thus avoided, which improves the description of the boundary interactions. A typical problem of these models is the slow or even absence of convergence as the array is expanded.

Stefanovich and Truong[36] stressed the need of well-chosen and optimized point charge arrays. The disadvantage of many methods[37−41] is that the optimal embedding charges are defined as the ones that reproduce a reference electrostatic potential which is assumed to be originating from a lattice consisting of ions bearing the formal ionic charges. This cannot be entirely justified, as all chemical bonds to some extent have a covalent component.

This obstacle can be overcome by optimizing the surrounding point charges in a manner that experimental data of the bulk material is reproduced by the cluster model. Xu et al.[31] varied the charges, placed at the genuine lattice positions, from 0.0 to ± 2.0 au in a study of CO adsorption on a MgO surface and determined an ideal value on the basis of the correspondence of calculated electronic and geometric properties with experimental ones. They observed indeed that electronic characteristics were strongly dependent on the value of the embedding point charges and that the nominal value of ±2.0 au overestimates the crystal potential. Their results underscore the need to draw a clear distinction between the formal charges of the ions constituting the real crystal and the embedding charges which best represent the rest of the infinite solid for a specific cluster.

We based our cluster model on a recent study of Kadossov et al.[29] The embedding point charges are located at the ideal lattice positions (the negative ones at the oxygen anion positions and the positive ones at the cationic sites), a choice that should lead to the best results.[40] Their values are determined by performing geometry optimizations and requiring a correspondence of the calculated lattice parameter with the experimental one. It is shown in ref 29 that this is a sufficient demand to ensure electronic properties being decently described.

3.2. Analysis of the Alkaline Earth Oxide Clusters. Before determination of the optimal values of the surrounding point charges, a well cut-out cluster needs to be constructed. As metal oxides are generally ionic species, their electronic structure has a rather localized nature, which implies that converged results with respect to cluster size are readily obtained with medium sized clusters.[31,42] We have therefore decided to use Me_9O_9 clusters (Me = Mg, Ca, Sr or Ba) with the atoms placed at the experimental rock salt structure positions in a 3 × 3 × 2 geometry, as can be seen in Figure 1a for the Sr_9O_9 case. Because we are focused on the reactive behavior of the surface O-atoms, the central oxygen in the upper layer constitutes the reactive site in this study. It is the only atom which is completely surrounded by quantum atoms, except for the surface side, of course, and consequently suffers the least from boundary effects. It is important

Table 1. Absolute Value of the Optimal Point Charges, q, Yielding the Experimental Lattice Parameter a_{exp} upon Geometry Optimization

system	q, au	a_{exp}, Å
MgO	0.427	4.21
CaO	0.516	4.81
SrO	0.444	5.11
BaO	0.510	5.52

to point out that the cluster matches the correct local C_{4v} symmetry of the reactive site which is essential for a realistic model.

The three principles mentioned in paragraph 3.1 are fulfilled. The stoichiometry and neutrality requirements are clearly satisfied, as is the coordination condition since the chosen geometry is the arrangement containing the most bonds.

The clusters are embedded in four layers of point charges at each side, apart from the surface one, which results in an $11 \times 11 \times 6$ point charge array at the genuine lattice positions. The authors of ref 29 for example analyzed the effect of larger embedding arrays and concluded that no significant improvement could be observed. Geometry optimizations for the clusters are performed for charge values varying form 0.0 to ± 2.0 au, and the ideal charges are identified as the ones giving rise to the experimental lattice parameters. We used the eigenvalue-following algorithm[43] as optimization technique with the lattice parameter as the only variable. All bonding angles were fixed at 90° so that the undistorted rock salt structure was assured. DFT calculations were done with the B3LYP approximation to the exchange-correlation functional.[44] The 6-311++G* basis set[45] was used for the O atoms, whereas we opted for the Stuttgard/Dresden pseudopotentials and corresponding basis sets for the metal atoms.[46] One should be cautious when selecting the number of core electrons to be replaced by a pseudopotential. As the metal atoms are in their cationic form, it is important to explicitly treat not only the two electrons from the outer shell, which are in fact partially removed by the surrounding oxygens, but also those from the lower shell. We therefore used the MWB10 pseudopotential and basis set for Ca, the MWB28 for Sr, and the MWB46 for Ba. For Mg, no pseudopotential satisfies the requirement and so the all-electron 6-311++G* basis set was used. All calculations were performed with the Gaussian 03 program.[47]

Table 1 summarizes the results of the point charge optimization. It seems that the embedding charges have ideal values of around $\pm$ 0.5 au, largely deviating from the formal charge values of ± 2.0 au. This clearly indicates that care should be taken when using formal charges in the construction of an embedding scheme, even for highly ionic systems. Incidentally, a formal charge embedding in our cluster model gives rise to an overestimation of the lattice parameter by roughly 50%, which indicates the very poor representation of the crystal potential with such an approach.

We conclude this section with a few notes on the electronic structure of the clusters in order to show the need for the MO-averaged Fukui function concept in the local reactivity description. As we want to investigate a reactivity trend of the surface oxygen anions, which are prone to electrophilic attacks, we only consider the (MO-averaged) Fukui function from below. The standard Fukui function $f^-(\mathbf{r})$ (eqs 13 and 15) gets its main contribution from the HOMO, which is given in Figure 1b for Sr_9O_9 and which has analogous characteristics in the other cases. The HOMO is, however, mainly located on the cluster edges and shows almost no density at the central surface oxygen. $f^-(\mathbf{r})$ will consequently not provide any significant reactivity information on the reactive site of interest but rather probes artificial edge effects. An analysis of the clusters' electronic structure holds the clue how $\tilde{f}^-(\mathbf{r})$ solves the problem.

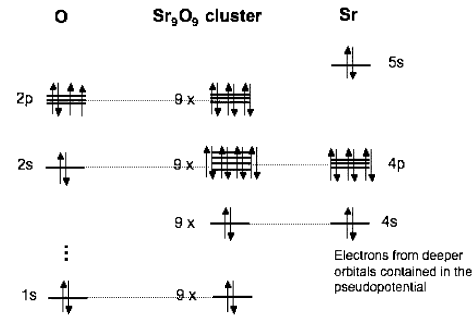


Figure 2. Schematic representation of the MO diagram of the Sr_9O_9 cluster.

Figure 2 gives a schematic representation of the MO diagram for the Sr_9O_9 cluster. Globally, the lowest 9 MOs are nonbonding orbitals with O-1s character; the next 9 MOs are nonbonding as well but now composed of the Sr-4s atomic orbitals. Then 36 MOs are found which are important contributors to the bonding of the cluster, as the O-2s and Sr-4p orbitals are mixed, whereas the final 27 occupied MOs are just located on the oxygens and are combinations of the O-2p atomic orbitals. The ionic character of the system manifests itself in the fact that the Sr-5s electrons move to the free O-2p places upon bonding. Most importantly, the 27 highest occupied MOs are linear combinations of the 2p atomic orbitals from the various O-atoms and thus all provide important contributions to the description of their reactivity and should be considered in the calculation of $\tilde{f}^-(\mathbf{r})$. Analogous conclusions can be drawn for the other systems. The calculation of the weighting factors w_i (eq 17) for these orbitals will be presented in the next section. The energetic criterion for MOs to enter the $\tilde{f}^-(\mathbf{r})$ expression is satisfied as the 27 highest orbitals are closely spaced (typically by 10^{-1} eV) and clearly energetically distinguishable from the others (with a typical gap of 10 eV).

4. Methodological Details and Practical Results

4.1. Calculation of the MO-Averaged Fukui Function. As mentioned in the theoretical section, the Fukui functions $f_i^-(\mathbf{r})$ needed in eq 17 are calculated as functional derivatives of the ith orbital energy ε_i with respect to $v(\mathbf{r})$:

$$f_i^-(\mathbf{r}) = \left(\frac{\delta \varepsilon_i}{\delta v(\mathbf{r})}\right)_N \tag{18}$$

Recently a straightforward scheme to evaluate these kind of derivatives was presented by the present authors in collaboration with P. W. Ayers.[27,48,49] We slightly improved the code for this study to include second-order effects and to take advantage of the molecular symmetry (vide infra). Details on the initial approach, its validation, and numerical results can be found in the literature.[27,48,49]

Consider a functional Taylor series expansion of the orbital energy ε_i as a functional of $v(\mathbf{r})$ up to second order:

$$\varepsilon_i[v(\mathbf{r}) + \Delta v(\mathbf{r})] = \varepsilon_i[v(\mathbf{r})] + \int \left(\frac{\delta \varepsilon_i}{\delta v(\mathbf{r})}\right)_N\Big|_{\Delta v(\mathbf{r})=0} \Delta v(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2}\iint \left(\frac{\delta^2 \varepsilon_i}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}\right)_N\Big|_{\Delta v(\mathbf{r})=0} \Delta v(\mathbf{r})\Delta v(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \quad (19)$$

Choosing the perturbations $\Delta v(\mathbf{r})$ in eq 19 of the external potential as $u_j(\mathbf{r})$ and $-u_j(\mathbf{r})$ and subtracting the corresponding expressions gives

$$\frac{1}{2}(\varepsilon_i[v(\mathbf{r}) + u_j(\mathbf{r})] - \varepsilon_i[v(\mathbf{r}) - u_j(\mathbf{r})]) = \int f_i^-(\mathbf{r})u_j(\mathbf{r})\mathrm{d}\mathbf{r} \quad (20)$$

The Fukui function $f_i^-(\mathbf{r})$ can now be found by expanding it in a basis set $\{\beta_k(\mathbf{r})\}_{k=1}^M$,

$$f_i^-(\mathbf{r}) = \sum_{k=1}^{M} c_k^i \beta_k(\mathbf{r}) \quad (21)$$

and solving the following set of linear equations for the unknown expansion coefficients c_k^i:

$$\frac{1}{2}(\varepsilon_i[v(\mathbf{r}) + u_j(\mathbf{r})] - \varepsilon_i[v(\mathbf{r}) - u_j(\mathbf{r})]) = \sum_{k=1}^{M} c_k^i \int \beta_k(\mathbf{r})u_j(\mathbf{r})\mathrm{d}\mathbf{r} \quad (22)$$

where the index j runs over 1 to P. As P is generally chosen to be much larger than M in order to provide an adequate description of the functional derivative, the set of eq 22 can be solved by a linear least-squares procedure.

The practical application of this approach requires a set of P perturbations $\{u_j(\mathbf{r})\}_{j=1}^P$ and their negative counterparts to be defined so as to calculate the left-hand side of eq 22. We will perturb the external potential by placing point charges in the molecular environment. For every positive point charge q_j the orbital energy $\varepsilon_i[v(\mathbf{r}) + u_j(\mathbf{r})]$ can be calculated by means of a single-point energy calculation as is the case for its negative counterpart $-q_j$ for which the corresponding orbital energy $\varepsilon_i[v(\mathbf{r}) - u_j(\mathbf{r})]$ is found. The integrals on the right-hand side of eq 22 are readily obtained as the perturbations $\{u_j(\mathbf{r})\}_{j=1}^P$ and the basis functions $\{\beta_k(\mathbf{r})\}_{k=1}^M$, for which s- and p-type Gaussians will be used, are known. Further details on the kind of basis functions and the characteristics of the point charges are given in paragraph 4.2.

We conclude this methodological section with a few notes on the weighting factors w_i in eq 17. As the central surface oxygen atom $O_{central}$ in our clusters (see Figure 1a) constitutes the reactive site of interest, the weighting factors should be proportional to the contribution of this atom to the various MOs entering eq 17. We will use a Mulliken population analysis[50] type of argument to give a rough estimate of this contribution. Since every molecular orbital $\psi_i(\mathbf{r})$ is expanded in a basis of atomic orbitals (AOs) $\phi_k(\mathbf{r})$, taken real for the sake of simplicity,

$$\psi_i(\mathbf{r}) = \sum_{k}^{\mathrm{AO}} d_k^i \phi_k(\mathbf{r}) \quad (23)$$

the share of $O_{central}$ in its integrated probability distribution can be estimated as $\sum_k^{O_{central}} \sum_l d_k^i d_l^i S_{kl}$, where S_{kl} is the overlap integral of two atomic orbitals k and l. A practical expression for w_i is obtained after normalization:

$$w_i = \left(\sum_k^{O_{central}} \sum_l d_k^i d_l^i s_{kl}\right) \Big/ \left(\sum_j \sum_k^{O_{central}} \sum_l d_i^i d_l^i s_{kl}\right) \quad (24)$$

where the summation over j includes all the MOs taken into account in eq 17.

4.2. Surface Reactivity Trend. The actual calculation of the functional derivatives requires the choice of a number of parameters. We refer to our previous studies[27,48,49] for a detailed description of the parameter optimization and immediately proceed to give the technical specifications used in the present study. The perturbations $\{\pm u_j(\mathbf{r})\}_{j=1}^P$ of the external potential are modeled as positive and negative point charges which are placed on a cubic grid within a region characterized by three scaled molecular van der Waals surfaces. Perturbations located between the surfaces scaled by the factors 0.05 and 0.75 are assigned charge values of ± 0.02 au, and those situated between the 0.75 and 2.50 scaled surfaces have charges of ± 0.05 au. The perturbations lying deeply in the molecular electron density only bear a small amount of charge so that the second-order approximation to the functional Taylor series expansion (eq 19) is justified.

Given the C_{4v} symmetry of the clusters under consideration, not the entire system but only one-eighth of it should be sampled by perturbations. Through the use of the molecular symmetry, these data can be multiplied throughout space to provide the necessary information for the calculation of the functional derivatives of the whole system (eq 22). We worked with a number of about 1500 perturbations, which totals a number of 12 000 perturbations after symmetry-based multiplication. The corresponding data set is largely sufficient to ensure well-converged results.

The basis set in which the functional derivatives are expanded (eq 21) is made up of the uncontracted s- and p-type functions of the Ahlrichs Coulomb fitting auxiliary basis set.[51] This density-based set is typically used for the acceleration of the evaluation of the Coulomb integrals in DFT calculations and has proven to be very successful in the representation of Fukui functions. The single-point calculations are carried out at the same level of theory as explained in section 3.2, except for the exchange-correlation functional that is replaced by the PBE approximation.[52]

Figures 1c and 1d give three-dimensional isosurface plots of the standard Fukui function $f^-(\mathbf{r})$ and the MO-averaged

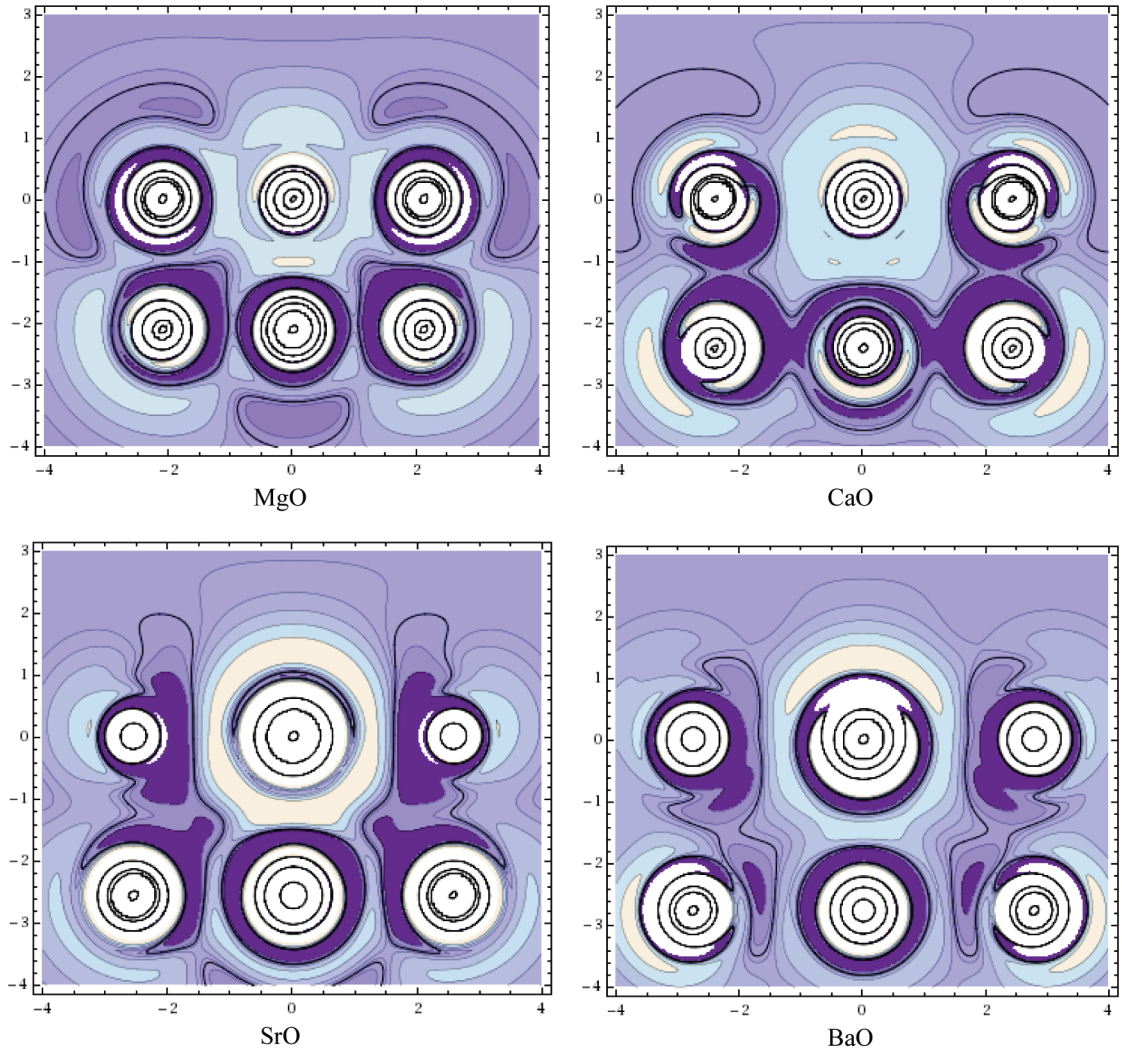


Figure 3. Contour plots of the MO-averaged Fukui function from below $\tilde{f}^{-}(\mathbf{r})$ for the Me_9O_9 clusters (with Me = Mg, Ca, Sr, or Ba), perpendicular to the clusters' surface. The contour values n_i (in au) are obtained as $n_i = 2 \times 10^{-5}\, i^3 - 3 \times 10^{-5}\, i^2 + 10^{-4}\, i$, with $i \in \{-2, -1, ..., 5\}$. The zero-contour is printed in bold and lighter regions are associated with higher values. The central surface oxygen atom lies at the (0 Å, 0 Å) coordinate.

function $\tilde{f}^{-}(\mathbf{r})$ for the Sr_9O_9 cluster. The standard Fukui function is predominantly located on the oxygens of the cluster edges and consequently does not provide a valuable description of the central oxygen, which is, as stated above, the best described quantum atom and should be used for the construction of an intermolecular reactivity scale. This erroneous behavior is directly linked to the electron distribution of the HOMO, which is not very pronounced at this central atom and is thus not the orbital that should be used to discuss the chemical reactivity. The introduction of the molecular orbitals made up of the O-2p atomic orbitals, as suggested in paragraph 3.2, in the calculation of the MO-averaged Fukui function $\tilde{f}^{-}(\mathbf{r})$ dramatically improves the results, however. Figure 1d gives the corresponding plot, which clearly shows how the central oxygen atom now governs the reactivity description of the cluster. As an analysis of the other systems (Mg_9O_9, Ca_9O_9, and Ba_9O_9) leads to analogous conclusions, the intercluster reactivity trend can now be constructed.

Figure 3 provides contour plots of the MO-averaged Fukui functions for the various Me_9O_9 systems. They are given in a plane perpendicular to the clusters' surface. Analysis of the contour lines at the central surface oxygen atom reveals that the values of the MO-averaged Fukui function increase as one goes down the periodic table, but somewhat stagnate between SrO and BaO, where no obvious difference can be noticed. This kind of result is of particular importance in the prediction and interpretation of a system's inherent reactivity toward electrophilic attacks, as the (MO-averaged) Fukui function is a measure of the change in electronic energy of the system under study when it undergoes a simultaneous change in external potential and in its number of electrons. A detailed discussion on this subject can be found in ref 53. The results can thus prove advantageous in the description of adsorption characteristics, catalytic activity, or reaction mechanisms. An increased catalytic activity trend from MgO to BaO has for example been found on the basis of experimental data for the destructive adsorption of carbon tetrachloride, a soft reagent which typically requests a soft reactivity index as the Fukui function in a theoretical analysis.[15]

The trend of increasing reactivity following the trends of increasing cation size and decreasing ionicity of the systems is extensively discussed in a previous study by some of the

present authors.[20] That study focused on the investigation of the local softness $s(\mathbf{r})$ of the alkaline earth oxide series through a periodic boundary conditions (PBC) approach and the use of the local density of states $g(r,E)$:[26,54]

$$s(\mathbf{r}) = \lim_{\delta\mu\to 0} \frac{1}{\delta\mu} \int_{\mu-\delta\mu}^{\mu} g(\mathbf{r}, E)\mathrm{d}E \qquad (25)$$

The fact that our results are also in agreement with their theoretical observations proves that our simple cluster model serves as a valuable and straightforward alternative to the infinite system approach. It is interesting to stress that our MO-averaged Fukui function concept occupies an intermediate place between the standard Fukui functions and the density of states-based descriptors. The former only use one frontier orbital in the description of a molecule's reactivity, whereas the latter utilize a band of orbitals to account for the reactivity of infinite systems. The use of an MO-averaged Fukui function $\tilde{f}(\mathbf{r})$ will typically prove to be beneficial in the intermediate situations where the valence orbitals are closely enough spaced to play an important role in the chemistry and yet do not form a continuous band.

5. Conclusion

Central in this paper is the reliable computation of a DFT-based reactivity index, the Fukui function, for cluster models of solid-state compounds. In the first part, we introduced the MO-averaged Fukui function concept $\tilde{f}^{-}(\mathbf{r})$. Its necessity comes from the observation that standard DFT reactivity descriptors assign a prominent role to the HOMO and the LUMO, but that these orbitals do not always satisfactorily represent the reactive site of interest. $\tilde{f}^{-}(\mathbf{r})$ bypasses this shortcoming by systematically placing some other, well-chosen MOs on the same footing.

The calculation of this new descriptor has been applied to the study of the alkaline earth oxide series. In order to avoid the PBC approach, we performed cluster-based calculations. It has been shown that a medium-sized quantum cluster embedded in an array of point charges is a sound model for the representation of this kind of systems, provided that the embedding charges are adequately optimized. The recovery of the experimental lattice parameter upon geometry optimization of the clusters has proven to be a straightforward and reliable way to obtain the required charges.

The final target of this contribution was the calculation of functional derivatives with respect to the external potential, for which an in-house method has been applied. Together, these three topics offer a valuable alternative to the density of states-based DFT concepts resulting from infinite system calculations. It has indeed been shown that for a series of alkaline earth oxides the results obtained in our work are completely in line with the PBC calculations invoking the local density of states to compute the Fukui function.

Acknowledgment. P.G. and F.D.P. acknowledge the Vrije Universiteit Brussel (VUB) and the Research Foundation - Flanders (FWO) for continuous support to their research group. N.S. acknowledges the FWO for a position as research assistant.

References

(1) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989; pp 1−333.

(2) Chermette, H. *J. Comput. Chem.* **1999**, *20*, 129–154.

(3) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1873, references therein.

(4) Geerlings, P.; De Proft, F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3028–3042.

(5) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. *J. Chem. Phys.* **1978**, *68*, 3801–3807.

(6) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516.

(7) Parr, R. G.; Yang, W. J. *J. Am. Chem. Soc.* **1984**, *106*, 4049–4050.

(8) Ayers, P. W.; Levy, M. *Theor. Chem. Acc.* **2000**, *103*, 353–360.

(9) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133–1138.

(10) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864–871.

(11) Yang, W.; Parr, R. G.; Pucci, R. *J. Chem. Phys.* **1984**, *81*, 2862–2863.

(12) Senet, P. *J. Chem. Phys.* **1997**, *107*, 2516–2524.

(13) Ayers, P. W. *Theor. Chem. Acc.* **2001**, *106*, 271–279.

(14) Soick, M.; Buyevskaya, O.; Höhenberger, M.; Wolf, B. *Catal. Today* **1996**, *32*, 163–169.

(15) Weckhuysen, B. M.; Mestl, G.; Rosyneck, M. P.; Krawitz, T. R.; Haw, J. F.; Lunsford, J. H. *J. Phys. Chem. B* **1998**, *102*, 3773–3778.

(16) Grönbeck, H.; Broqvist, P.; Panas, I. *Surf. Sci.* **2006**, *600*, 403–408.

(17) Grönbeck, H.; Broqvist, P. *J. Chem. Phys.* **2003**, *119*, 3896–3904.

(18) Zecchina, A.; Scarano, D.; Bordiga, S.; Spoto, G.; Lamberti, C. *Adv. Catal.* **2001**, *46*, 265–397.

(19) Ono, Y.; Baba, T. *Catal. Today* **1997**, *38*, 321–337.

(20) Cárdenas, C.; De Proft, F.; Chamorro, E.; Fuentealba, P.; Geerlings, P. *J. Chem. Phys.* **2008**, *128*, 034708.

(21) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.

(22) Zhang, Y.; Yang, W. *Theor. Chem. Acc.* **2000**, *103*, 346–348.

(23) Ayers, P. W. *J. Math. Chem.* **2008**, *43*, 285–303.

(24) Janak, J. F. *Phys. Rev. B* **1978**, *18*, 7165–7168.

(25) Chan, G. K.-L. *J. Chem. Phys.* **1999**, *110*, 4710–4723.

(26) Cohen, M. H.; Ganduglia-Pirovano, M. V. *J. Chem. Phys.* **1994**, *101*, 8988–8997.

(27) Fievez, T.; Sablon, N.; De Proft, F.; Ayers, P. W.; Geerlings, P. *J. Chem. Theory Comput.* **2008**, *4*, 1065–1072.

(28) Sauer, J. *Chem. Rev.* **1989**, *89*, 199–255.

(29) Kadossov, E. B.; Gaskell, K. J.; Langell, M. A. *J. Comput. Chem.* **2007**, *28*, 1240–1251.

(30) Stefanovich, E. V.; Truong, T. N. *J. Chem. Phys.* **1996**, *104*, 2946–2955.

(31) Xu, X.; Nakatsuji, H.; Lu, X.; Ehara, M.; Cai, Y.; Wang, N. Q.; Zhang, Q. E. *Theor. Chem. Acc.* **1999**, *102*, 170–179.

(32) Lü, X.; Xu, X.; Wang, N.; Zhang, Q.; Ehara, M.; Nakatsuji, H. *Chem. Phys. Lett.* **1998**, *291*, 445–452.

(33) Xu, X.; Lü, X.; Wang, N.; Zhang, Q. *Chem. Phys. Lett.* **1995**, *235*, 541–547.

(34) (a) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733. (b) Masera, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(35) (a) Vail, J. M. *J. Phys. Chem. Solids* **1990**, *51*, 589–607. (b) Yudanov, I. V.; Nasluzov, V. A.; Neyman, K. M.; Rösch, N. *Int. J. Quantum Chem.* **1997**, *65*, 975–986.

(36) Stefanovich, E. V.; Truong, T. N. *J. Phys. Chem. B* **1998**, *102*, 3018–3022.

(37) Sousa, C.; Casanovas, J.; Rubio, J.; Illas, F. *J. Comput. Chem.* **1993**, *14*, 680–684.

(38) Huang, Z. H.; Guo, H. *Surf. Sci.* **1993**, *286*, 182–189.

(39) Rittner, F.; Fink, R.; Boddenberg, B.; Staemmler, V. *Phys. Rev. B* **1999**, *57*, 4160–4171.

(40) Derenzo, S. E.; Klintenberg, M. K.; Weber, M. J. *J. Chem. Phys.* **2000**, *112*, 2074–2081.

(41) Batista, E. R.; Friesner, R. A. *J. Phys. Chem. B* **2002**, *106*, 8136–8141.

(42) Xu, X.; Nakatsuji, H.; Ehara, M.; Lü, X.; Wang, N. Q.; Zhang, Q. E. *Chem. Phys. Lett.* **1998**, *292*, 282–288.

(43) (a) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800–2806. (b) Simons, S.; Jorgensen, P.; Taylor, H.; Ozment, J. *J. Phys. Chem.* **1983**, *87*, 2745–2753. (c) Bannerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52–57.

(44) (a) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789. (b) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (c) Stevens, P. J.; Delvin, F. J.; Chablaoski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(45) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986; Chapter 4; pp 65−88.

(46) Kaupp, M.; Schleyer, P. v. R.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1991**, *94*, 1360–1366.

(47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery,Jr. J. A.; 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.; Kno 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 D.01, Gaussian, Inc., Wallingford, CT, 2005.

(48) Ayers, P. W.; De Proft, F.; Borgoo, A.; Geerlings, P. *J. Chem. Phys.* **2007**, *126*, 224107.

(49) Sablon, N.; Ayers, P. W.; De Proft, F.; Geerlings, P. *J. Chem. Phys.* **2007**, *126*, 224108.

(50) Mulliken, R. S. *J. Chem. Phys.* **1934**, *2*, 782–793.

(51) (a) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289. (b) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(52) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(53) Ayers, P. W.; Anderson, J. S. M.; Bartolotti, L. J. *Int. J. Quantum Chem.* **2005**, *101*, 520–534.

(54) (a) Brommer, K. D.; Galván, M.; Dal Pino, A.; Joannopoulos, J. D. *Surf. Sci.* **1994**, *314*, 57–70. (b) Nguyen, L. T.; De Proft, F.; Amat, M. C.; Van Lier, G.; Fowler, P. W.; Geerlings, P. *J. Phys. Chem. A* **2003**, *107*, 6837–6842.

CT9000312

JCTC Journal of Chemical Theory and Computation

Energies, Geometries, and Charge Distributions of Zn Molecules, Clusters, and Biocenters from Coupled Cluster, Density Functional, and Neglect of Diatomic Differential Overlap Models

Anastassia Sorkin,[†] Donald G. Truhlar,[‡] and Elizabeth A. Amin[†,*]

Department of Medicinal Chemistry, College of Pharmacy, University of Minnesota, 717 Delaware St. SE, Minneapolis, Minnesota 55414-2959, and Department of, Chemistry, University of Minnesota, 207 Pleasant St. SE, Minneapolis, Minnesota, 55455-0431

Received January 20, 2009

Abstract: We present benchmark databases of Zn−ligand bond distances, bond angles, dipole moments, and bond dissociation energies for Zn-containing small molecules and Zn coordination compounds with H, CH_3, C_2H_5, NH_3, O, OH, H_2O, F, Cl, S, and SCH_3 ligands. The test set also includes clusters with Zn−Zn bonds. In addition, we calculated dipole moments and binding energies for Zn centers in coordination environments taken from zinc metalloenzyme X-ray structures, representing both structural and catalytic zinc centers. The benchmark values are based on relativistic-core coupled cluster calculations. These benchmark calculations are used to test the predictions of four density functionals, namely B3LYP and the more recently developed M05-2X, M06, and M06-2X levels of theory, and six semiempirical methods, including neglect of diatomic differential overlap (NDDO) calculations incorporating the new PM3 parameter set for Zn called ZnB, developed by Brothers and co-workers, and the recent PM6 parametrization of Stewart. We found that the best DFT method to reproduce dipole moments and dissociation energies of our Zn compound database is M05-2X, which is consistent with a previous study employing a much smaller and less diverse database and a much larger set of density functionals. Here we show that M05-2X geometries and single-point coupled cluster calculations with M05-2X geometries can also be used as benchmarks for larger compounds, where coupled cluster optimization is impractical, and in particular we use this strategy to extend the geometry, binding energy, and dipole moment databases to additional molecules, and we extend the tests involving crystal-site coordination compounds to two additional proteins. We find that the most predictive NDDO methods for our training set are PM3 and MNDO/d. Notably, we also find large errors in B3LYP for the coordination compounds based on experimental X-ray geometries.

1. Introduction

Zn compounds attract considerable interest in the fields of petroleum chemistry,[1−10] sensing,[5−14] nanotechnology,[9−11,15−34] and bioinorganic chemistry;[35−76] zinc metalloproteins in particular play crucial roles in various biochemical processes and are important drug targets. The reliability of density functional theory (DFT) and neglect of diatomic differential overlap (NDDO) semiempirical molecular orbital calculations for modeling and predicting geometric, electronic, and energetic features of zinc compounds is not well established, although several attempts to evaluate the accuracy of these methods have been published.[77−91] Benchmark databases containing Zn−ligand bond distances, dipole moments, and dissociation energies in Zn

* Corresponding author e-mail: eamin@umn.edu.
[†] Department of Medicinal Chemistry.
[‡] Department of Chemistry.

10.1021/ct900038m CCC: $40.75
Published on Web 04/02/2009

Table 1. CCSD(T) Best Estimates of Zn−Ligand Bond Distances (in Å) for 26 Zn Model Compounds (from database D1), with Corresponding DFT and CCSD Values and Mean Unsigned Errors (MUEs)

compd	bond	B3LYP	M05-2X	M06	M06-2X	CCSD	CCSD(T)
ZnH^+	Zn−H	1.520	1.533	1.517	1.548	1.524	1.526
ZnH	Zn−H	1.624	1.642	1.640	1.646	1.605	1.605
ZnH_2	Zn−H	1.530	1.550	1.529	1.562	1.531	1.528[a]
MUE	Zn−H	0.009	0.022	0.015	0.032	0.002	
$ZnMe^+$	Zn−C	1.990	1.992	1.957	2.005	1.972	1.976
ZnMe	Zn−C	2.056	2.067	2.066	2.056	2.014	2.014
$ZnCN^+$	Zn−C	1.857	1.866	1.849	1.880	1.852	1.852
ZnCN	Zn−C	1.970	1.985	1.973	1.998	1.955	1.951
MUE	Zn−C	0.020	0.029	0.024	0.037	0.002	
$Zn(NH_3)^{2+}$	Zn−N	1.947	1.948	1.919	1.956	1.934	1.939[a]
$Zn(NH_3)_3^{2+}$	Zn−N	2.019	2.020	1.994	2.031	2.001	2.005[a,b]
$Zn(NH_3)_4^{2+}$	Zn−N	2.082	2.076	2.054	2.089	2.058	2.063[a,b]
MUE	Zn−N	0.014	0.013	0.014	0.023	0.005	
ZnO(singlet)	Zn−O	1.704	1.707	1.694	1.726	1.715	1.711[a]
ZnO(triplet)	Zn−O	1.885	1.890	1.885	1.891	1.852	1.855
$Zn(OH)^+$	Zn−O	1.762	1.751	1.740	1.764	1.748	1.757[a]
$Zn(OH)_2$	Zn−O	1.767	1.771	1.750	1.781	1.758	1.767[a]
$Zn(H_2O)^{2+}$	Zn−O	1.857	1.860	1.843	1.867	1.853	1.852[a]
MUE	Zn−O	0.009	0.011	0.018	0.017	0.005	
ZnS	Zn−S	2.067	2.079	2.050	2.096	2.068	2.066[a]
$ZnSH^+$	Zn−S	2.157	2.159	2.135	2.171	2.148	2.149
$Zn(SCH_3)^+$	Zn−S	2.178	2.175	2.153	2.190	2.161	2.170[a]
MUE	Zn−S	0.006	0.008	0.017	0.024	0.004	
ZnF	Zn−F	1.797	1.783	1.791	1.794	1.773	1.775
ZnF_2	Zn−F	1.730	1.731	1.717	1.739	1.722	1.723
$ZnCl^+$	Zn−Cl	2.060	2.064	2.034	2.071	2.050	2.053
ZnCl	Zn−Cl	2.179	2.181	2.159	2.185	2.152	2.151
$ZnCl_2$	Zn−Cl	2.087	2.098	2.066	2.106	2.080	2.078
MUE	Zn−halogen	0.015	0.015	0.012	0.023	0.002	
Zn_2	Zn−Zn	4.080	3.900	3.306	4.071	4.454	4.104
Zn_3	Zn−Zn	3.550	3.544	3.132	4.045	4.220	3.674
Zn_4	Zn−Zn	3.016	3.097	2.906	3.538	3.081	2.802
MUE	Zn−Zn	0.121	0.210	0.481	0.380	0.391	
MUE[c]		0.013	0.015	0.017	0.025	0.003	
MUE[d]		0.026	0.041	0.075	0.070	0.052	
MUE[e]		0.025	0.039	0.070	0.066	0.048	

[a] Values published in ref 89. [b] Values estimated by extrapolation from MP2/B2. [c] MUE without Zn−Zn bonds (23 bond lengths). [d] MUE without extrapolated values (24 bond lengths). [e] MUE over all bond distances (26 bond lengths).

model compounds with H, O, OH, H_2O, NH_3, S, and SCH_3 were presented recently,[89] together with an evaluation of various density functional approximations. In that study, out of 39 density functionals and eight other molecular orbital methods that were tested, the M05-2X[92] density functional was recommended as the best affordable method to obtain accurate geometric parameters, dipole moments, and bond dissociation energies for the selected Zn centers.

In the present paper, we greatly enlarge the benchmark database for Zn compounds presented in ref 89 in terms of the number of compounds and the diversity of coordination motifs. Our broadened test set has two parts.

The first part represents Zn molecules and clusters and it includes compounds with Zn−H, Zn−C, Zn−N, Zn−O, Zn−S, and Zn−halogen bonds and clusters with two to four Zn atoms; this part comprises two new databases. The first (called database D1), for molecules and clusters, contains 26 bond distances, nine bond angles, 20 dipole moments, and 29 dissociation energies. As experimental values for these parameters are not currently available, benchmark geometries and bond dissociation energies in this database are obtained by relativistic-core coupled cluster theory with single and double excitations and a quasiperturbative treatment of connected triple excitations [CCSD(T)[93]]; dipole moments are obtained from relativistic-core CCSD[94] calculations. Including relativistic-core M05-2X geometries and relativistic-core coupled cluster calculations at relativistic-core M05-2X geometries as benchmarks permits us to enlarge the database to 43 bond distances, 38 bond angles, 22 dipole moments, and 36 dissociation energies (database D2).

The geometries obtained with four density functionals (B3LYP,[95−98] M05-2X,[92] M06,[99,100] and M06-2X[99,100]) are tested against CCSD(T) benchmarks, while the geometries produced by NDDO methods [AM1,[101−104] PM3,[105] PM3-(tm),[105−109] MNDO/d,[110,111] PM6[112] and the new parametrization of PM3 method for Zn named ZnB recently proposed by Brothers et al.[84]] are evaluated against the larger D2 benchmark database incorporating the M05-2X bond distances and bond angles. Dipole moments and dissociation energies from DFT and NDDO calculations are also tested against this enlarged benchmark database.

The second part of our test set contains four zinc coordination compounds, three inorganic and one organometallic, designed to mimic zinc sites in metalloproteins. For these structures, we fix key geometric parameters at their experimental metalloenzyme X-ray structure values and use these structures to test approximate methods for dissociation energies, bond stretching energies, dipole moments, and

Table 2. M05-2X Benchmark Bond Distances (in Å) for 17 Additional Zn Model Compounds

compd	bond	M05-2X
$ZnMe_2$	Zn−C	1.962
ZnEt	Zn−C	2.108
$ZnEt_2$	Zn−C	1.977
$Zn(CN)_2$	Zn−C	1.894
$Zn(NH_3)_2^{2+}$	Zn−N	1.945
$Zn(NH_3)_2(OH)_2$	Zn−N	2.179
$Zn(NH_3)(OH)^+$	Zn−N	1.954
$Zn(NH_3)(OH)_2$	Zn−N	2.158
Zn_2O_2	Zn−O	1.869
Zn_3O_3	Zn−O	1.813
Zn_4O_4	Zn−O	1.787
$Zn(NH_3)_2(OH)_2$	Zn−O	1.868
$Zn(NH_3)(OH)^+$	Zn−O	1.738
$Zn(NH_3)(OH)_2$	Zn−O	1.813
Zn_2S_2	Zn−S	2.255
Zn_3S_3	Zn−S	2.196
Zn_4S_4	Zn−S	2.176

dipole moments of zinc-containing fragments in which one bond is broken.

2. Data Sets and Computational Details

2.1. Methods and Basis Set. Four density functionals are assessed in this paper, namely, B3LYP,[95−98] M05-2X,[92] M06,[99,100] and M06-2X.[99,100] B3LYP is a hybrid generalized gradient approximation functional with 20% Hartree−Fock exchange. M05-2X, M06, and M06-2X are hybrid meta-functionals in which the kinetic energy density and improved functional forms are employed in the exchange functional, which in turn allows the correlation functional, which also depends on the kinetic energy density and involves new functional forms, to become more physical.[92,99,100] These functionals contain respectively 56%, 27%, and 54% Hartree−Fock exchange.

The less accurate but faster semiempirical methods that we tested are AM1,[101−104] PM3,[105] PM3(tm),[105−109] MNDO/d,[110,111] PM6,[112] and the new parametrization[84] of the PM3 method for Zn named ZnB. These methods are all based on the neglect of diatomic differential overlap (NDDO) approximation.

PM3 and AM1 are the two most widely used NDDO methods, and their suitability for organic and Zn coordination compounds has been reviewed recently.[79,113] MNDO/d is based on the earlier MNDO[114,115] method but with d-orbitals added for some elements. Due to its closed d-shell, zinc is treated only with s- and p-orbitals in AM1 and PM3.[111] A recent test[79] of AM1, PM3, and MNDO/d for zinc compounds showed that MNDO/d appears to be the most accurate method for the representation of bio-organic complexes containing Zn^{2+}, but it exhibits deficiencies in the description of Zn−S interactions. In the same study, PM3 failed to satisfactorily describe Zn−O interactions, but it proved to be the best method for the description of Zn−N complexes.[79]

PM3(tm) is an extension of the standard PM3 Hamiltonian for transition-metal elements, but published applications and evaluations are rare. According to ref 108, PM3(tm) is quite successful in predicting the geometries of transition-metal complexes. PM6 is a very recent method; it has been parametrized against a large training set.[112] The new ZnB (Zinc Biological) parameter set of PM3 was proposed by Brothers et al.[84] to improve the ability of PM3 to model biological zinc coordination sites.

Table 3. CCSD(T) Benchmark Values (database D1) for Nine Zn−Ligand Bond Angles (in deg), with Corresponding DFT/B2 Values and Mean Unsigned Errors (MUEs)

compd	angle	B3LYP	M05-2X	M06	M06-2X	CCSD(T)
$Zn(OH)^+$	Zn−O−H	112.31	115.39	112.82	114.51	111.89
$Zn(OH)_2$	O−Zn−O	174.36	176.23	174.78	176.15	175.57
$Zn(OH)_2$	Zn−O−H	114.42	116.99	114.17	116.52	115.18
$Zn(H_2O)^{2+}$	Zn−O−H	124.80	125.36	125.03	125.36	125.58
$ZnMe^+$	Zn−C−H	91.72	90.83	90.75	90.99	91.14
ZnMe	Zn−C−H	105.72	103.77	104.03	104.40	104.25
$ZnSH^+$	Zn−S−H	111.04	111.68	111.64	111.74	111.97
$Zn(SCH_3)^+$	Zn−S−C	103.56	104.16	104.05	104.07	104.01
$Zn(NH_3)^{2+}$	Zn−N−H	107.71	107.33	106.93	108.82	108.74
MUE		0.85	0.98	0.68	0.60	

Note that all NDDO methods evaluated here involve a "very large" core (28 electrons for AM1, PM3, PM6, and ZnB) or a "large" core [18 electrons for PM3(tm) and MNDO/d] for Zn, whereas the density functional and coupled cluster calculations involve a "small" core (10 electrons). For DFT and coupled cluster calculations, we use a relativistic effective core potential (ECP) and a previously[89] designed basis set for Zn, referred to as B2. In this basis, Zn is represented by a supplemented version of the basis set called 6-311+G(d,p)[116−118] in Gaussian 03,[119] which is constructed from the earlier work of Wachters[107] and Hay[117] and further augmented by two additional f functions specified by Raghavachari and Trucks,[118] with exponents of 0.486 and 5.40. The ECP used here is the MEFIT,*R* pseudopotential[120] of Preuss et al., which replaces the 10 innermost electrons.

For H, C, N, O, and F, the NDDO calculations use a two-electron core, and for S and Cl they use a 10-electron core, whereas the DFT and coupled cluster calculations treat all electrons explicitly for these elements. For H, C, N, O, F, S, and Cl, the DFT and coupled cluster calculations are carried out with all electrons explicit and with the MG3S[121] basis set.

We have recently added the Zn B2 basis set to the EMSL Basis Set Exchange,[122] where it is listed as "B2 basis set for Zn".

2.2. Software and Hardware. CCSD, CCSD(T), and DFT calculations in this paper were carried out on the Minnesota Supercomputing Institute core resources using MN-GFM version 3.0,[123] which is a locally modified version of Gaussian 03.[119] AM1, PM3, PM6, and ZnB calculations were performed using MOPAC2007[124] and MOPAC 5.013mn[125] on a Dell OptiPlex 745 machine running under SUSe Linux 10.1. MNDO/d and PM3(tm) calculations were carried out on the Dell OptiPlex 745 using SPARTAN'04 Linux.[126]

3. Benchmark Databases

3.1. Molecules and Clusters. For evaluation of density functionals and NDDO methods for Zn−ligand compounds, Zn_n clusters, Zn_nO_n clusters, and Zn_nS_n clusters, we used two

Table 4. M05-2X Benchmark Values for 29 Additional Zn−Ligand Bond Angles (in deg)

compd	angle	M05-2X
Zn_2O_2	O−Zn−O	102.17
Zn_3O_3	O−Zn−O	142.92
Zn_4O_4	O−Zn−O	162.82
Zn_2S_2	S−Zn−S	113.44
Zn_3S_3	S−Zn−S	156.35
Zn_4S_4	S−Zn−S	177.00
$ZnMe_2$	Zn−C−H	110.82
ZnEt	Zn−C−C	111.29
ZnEt	Zn−C−H	103.98
$ZnEt_2$	C−Zn−C	179.49
$ZnEt_2$	Zn−C−C	114.25
$ZnEt_2$	Zn−C−H	108.10
$Zn(NH_3)_2^{2+}$	Zn−N−H	112.53
$Zn(NH_3)_3^{2+}$	Zn−N−H	112.76
$Zn(NH_3)_4^{2+}$	Zn−N−H	113.11
$Zn(NH_3)(OH)^+$	N−Zn−O	171.77
$Zn(NH_3)(OH)^+$	Zn−N−H	112.48
$Zn(NH_3)(OH)^+$	Zn−O−H	120.84
$Zn(NH_3)_2(OH)_2$	N−Zn−O	115.91
$Zn(NH_3)_2(OH)_2$	O−Zn−O	145.05
$Zn(NH_3)_2(OH)_2$	N−Zn−N	104.59
$Zn(NH_3)_2(OH)_2$	Zn−O−H	119.55
$Zn(NH_3)_2(OH)_2$	Zn−N−H	91.76
$Zn(NH_3)_2(OH)_2$	Zn−N−H	114.09
$Zn(NH_3)_2(OH)_2$	Zn−N−H	121.71
$Zn(NH_3)(OH)_2$	N−Zn−O	93.24
$Zn(NH_3)(OH)_2$	O−Zn−O	161.49
$Zn(NH_3)(OH)_2$	Zn−O−H	117.34
$Zn(NH_3)(OH)_2$	Zn−N−H	113.66

Table 5. CCSD Dipole Moment Benchmark Values (relative to center of mass, in D) for 29 Zn Model Compounds (database D1), with Corresponding DFT and CCSD//M05-2X Values and Mean Unsigned Errors (MUEs)

	B3LYP	M05-2X	M06	M06-2X	CCSD//M05-2X	CCSD
ZnO(singlet)	5.53	6.28	5.34	5.98	5.50	5.50[a]
Zn−O(triplet)	2.43	2.37	2.60	2.53	2.73	2.62
Zn−S	5.09	2.67	4.68	5.58	5.48	5.47[a]
$Zn(OH)^+$	3.59	4.37	3.81	4.25	4.24	4.27[a]
$Zn(H_2O)^{2+}$	0.79	0.51	0.70	0.53	0.39	0.39[a]
$Zn(OH)_2$	1.87	1.71	1.90	1.64	1.65	1.72
$ZnSH^+$	3.06	3.75	3.08	3.80	3.79	3.78
$Zn(SCH_3)^+$	2.74	3.53	2.73	2.19	3.58	3.59[a]
ZnH^+	1.20	1.57	1.52	1.55	1.57	1.59
ZnH	0.49	0.35	0.55	0.50	0.64	0.57
$Zn(NH_3)^{2+}$	1.82	1.36	1.67	1.40	1.20	1.27[a]
$Zn(NH_3)(OH)^+$	7.05	7.35	7.09	7.33	7.27	7.34[a]
$Zn(NH_3)_2(OH)_2$	3.45	3.75	3.44	3.69	3.79	3.80
$ZnMe^+$	0.52	1.07	0.66	1.02	1.12	1.17
ZnMe	0.25	0.45	0.25	0.10	0.06	0.07
$ZnCN^+$	6.61	7.59	6.78	7.48	7.62	7.56
ZnCN	4.50	4.54	4.69	4.61	4.69	4.54
ZnF	3.17	3.18	3.43	3.23	3.25	3.20
$ZnCl^+$	4.78	5.47	4.84	5.47	5.58	5.56
ZnCl	2.84	2.74	2.86	2.88	3.07	2.92
MUE	0.39	0.14	0.36	0.16	0.05	

[a] Values published in ref 89.

benchmark databases. One of these (D1) contains CCSD(T) bond distances, bond angles, and bond dissociation energies, and CCSD dipole moments. This database is used for evaluating both density functional and semiempirical methods. The second database (D2) includes all data in D1 plus some M05-2X values for bond distances and bond angles, some CCSD values of dipole moments calculated in single-point calculations at M05-2X geometries, and some CCSD(T) values of dissociation energies calculated in single-point calculations at M05-2X geometries. The D2 benchmark geometries are used for evaluation of geometries calculated with NDDO methods, while the D2 dipole moments and dissociation energies are used to test both density functional and NDDO dipole moments and dissociation energies.

Table 6. Additional CCSD//M05-2X Benchmark Values for Two Zn Model Compound Dipole Moments (in D)

compd	CCSD//M05-2X
ZnEt	0.46
$Zn(NH_3)(OH)_2$	2.34

The CCSD(T) method has been shown to be reliable and accurate for reproducing experimental parameters for small transition-metal compounds.[127−130] Table 1 gives the best estimates of 26 Zn−ligand bond distances in database D1; some of these values were published previously in ref 89. The best estimates of $Zn(NH_3)_3^{2+}$ and $Zn(NH_3)_4^{2+}$ geometric parameters were obtained by extrapolation[89] from MP2/B2 calculations, while all other values in this table are from full CCSD(T)/B2 optimization.

Table 1 shows that the best DFT methods for Zn−ligand geometries are M05-2X and B3LYP; these methods can therefore be used to obtain additional benchmark values to evaluate the performance of less accurate semiempirical calculations for geometric parameters. We enlarged the benchmark database used to assess the accuracy of NDDO methods by adding 17 more M05-2X bond distances (Table 2) plus, as discussed below, 29 more bond angles, seven more dissociation energies, and two more dipole moments. All these additional values are included in the D2 database.

Table 3 gives the best estimates of nine Zn-ligand bond angles yielded by CCSD(T) calculations (database D1). The 29 M05-2X additional bond angle values included in the database D2 are found in Table 4.

The CCSD method yields very accurate geometries (Table 1), and CCSD optimizations are used to calculate best estimates of dipole moments. Table 5 lists CCSD dipole moments for 20 Zn−ligand compounds (database D1). The mean unsigned deviation of CCSD//M05-2X from CCSD is much smaller than the deviation of any of the four DFT methods from CCSD; therefore, CCSD//M05-2X calculations can be used to obtain suitable benchmarks to estimate the accuracy of both DFT and NDDO methods. Table 6 lists two dipole moments obtained by the CCSD//M05-2X method that are included in database D2.

Best estimates of dissociation energies were calculated by the CCSD(T) method. Table 7 lists DFT and CCSD(T)//M05-2X bond dissociation energies (data set D1). The CCSD(T)//M05-2X method gives results very close to full CCSD(T) values; once again, CCSD(T)//M05-2X can justifiably be used to calculate best estimates for systems where full coupled-cluster optimizations are not feasible. These additional benchmark values (included in database D2) are given in Table 8.

3.2. Coordination Compounds Based on X-ray Structures. To assess the performance of density functional approximations and NDDO techniques for structures that

Table 7. CCSD(T) Benchmarks for Bond Dissociation Energies (BDEs, kcal/mol) for 29 Zn-Containing Small Molecules (database D1), with CCSD(T)/M05-2X and DFT Values, and Mean Unsigned Errors (MUEs)

compd	products	B3LYP	M05-2X	M06	M06-2X	CCSD(T)//M05-2X	CCSD(T)
ZnO(singlet)	Zn, O(singlet)	90.02	76.66	81.23	81.38	80.32	80.32[a]
ZnO(singlet)	Zn, O(triplet)	26.62	19.77	20.04	21.60	29.14	29.14
ZnO(triplet)	Zn, O(singlet)	91.14	80.59	82.31	88.87	75.18	75.37
ZnO(triplet)	Zn, O(triplet)	27.73	23.70	21.11	29.09	23.99	24.18
ZnS	Zn, S(singlet)	64.82	66.95	61.28	66.30	58.54	58.56[a]
ZnS	Zn, S(triplet)	26.41	30.20	25.71	30.24	27.84	27.87
Zn_2	Zn, Zn	−0.05	0.60	1.90	0.52	0.41	0.43
Zn_3	Zn_2, Zn	0.17	2.16	5.01	1.75	1.22	1.22
Zn_4	Zn_3, Zn	2.84	6.69	10.88	4.43	6.43	8.00
ZnH_2	ZnH, H	79.11	80.70	80.41	75.46	78.88	78.90[a]
$Zn(OH)^+$	Zn^{2+}, OH^-	446.96	436.57	443.12	437.76	435.56	435.67[a]
$Zn(OH)_2$	$Zn(OH)^+$, OH^-	257.03	262.42	261.24	260.45	260.00	258.71[a]
$Zn(H_2O)^{2+}$	Zn^{2+}, H_2O	106.89	104.12	104.12	104.55	99.79	99.78
$ZnSH^+$	Zn^{2+}, SH^-	434.53	420.75	434.63	421.23	417.50	417.51
$Zn(SCH_3)^+$	Zn^{2+}, SCH_3^-	448.80	433.69	446.97	433.36	430.42	433.84[a]
$Zn(NH_3)^{2+}$	Zn^{2+}, NH_3	143.20	136.51	139.99	136.74	133.12	134.15[a]
$Zn(NH_3)_3^{2+}$	$Zn(NH_3)_2^{2+}$, NH_3	58.91	64.31	61.30	64.02	62.07	60.67[a,b]
$Zn(NH_3)_4^{2+}$	$Zn(NH_3)_3^{2+}$, NH_3	43.24	49.00	46.70	49.23		46.06[a,b]
$Zn(NH_3)(OH)^+$	$Zn(OH)^+$ NH_3	76.97	79.77	80.67	79.11	78.87	81.14[a,b]
$Zn(NH_3)_2(OH)_2$	$Zn(NH_3)(OH)_2$, NH_3	8.76	15.63	12.12	16.32	12.95	8.87[a,b]
$ZnMe^+$	Zn^{2+}, Me	496.87	483.52	494.47	482.78	480.73	480.77
ZnMe	Zn, Me	11.28	12.39	9.02	16.57	13.30	13.60
$ZnCN^+$	Zn^{2+}, CN^-	398.88	383.68	397.42	385.97	385.85	386.52
ZnCN	Zn, CN	50.08	53.76	42.90	57.08	54.24	54.87
ZnF	Zn, F	63.14	61.50	57.85	66.09	63.88	63.90
ZnF_2	ZnF,F	114.64	123.79	118.14	115.75	116.75	116.77
$ZnCl^+$	Zn^{2+}, Cl^-	404.23	393.20	404.16	393.77	390.01	390.04
ZnCl	Zn, Cl	43.36	45.40	40.49	51.20	45.77	45.88
$ZnCl_2$	ZnCl,Cl	95.24	105.85	101.85	100.42	99.08	99.11
MUE		6.04	3.13	5.54	3.36	0.63	

[a] Values published in ref 89. [b] Values published in ref 89 and estimated by extrapolation from MP2/B289.

Table 8. Additional CCSD(T)//M05-2X Benchmark Values for Bond Dissociation Energies (in kcal/mol) for Seven Zn-Containing Compounds

compd	products	CCSD(T)//M05-2X
Zn_2O_2	2 ZnO (singlet)	94.00
Zn_2O_2	2 ZnO (triplet)	104.29
Zn_2S_2	2 ZnS	93.97
$Zn(NH_3)_2^{2+}$	$Zn(NH_3)^{2+}$, NH_3	109.08
$ZnMe_2$	ZnMe, Me	73.16
ZnEt	Zn, Et	8.98
$ZnCN_2$	ZnCN, CN	115.86

more closely represent Zn biocenters, we created four model compounds based on Zn metalloprotein centers: anthrax toxin lethal factor, abbreviated LF (PDB ID 1PWU);[73] matrix metalloproteinase-1, abbreviated MMP-1, also known as collagenase-1 (PDB ID 2CLT);[74] matrix metalloproteinase-3, abbreviated MMP-3, also known as stromelysin-1, (PDB ID 1SLN);[75] and *Saccharomyces cerevisiae* YNR046W, a Zn-finger protein from the ERF1 methyltransferase complex (PDB ID 2J6A)[76] (Figure 1). Inclusion of these structures enables the accuracy of DFT and NDDO methods to be tested at geometries that are relevant to metalloprotein dynamics. To increase the diversity of our test set, we chose two pentacoordinate catalytic Zn centers (LF and MMP-3) and two tetracoordinate structural Zn centers (MMP-1 and the Zn-finger). In the anthrax toxin lethal factor enzyme, the catalytic Zn is ligated by two histidines and one glutamic acid; in 1PWU, this zinc is further coordinated by two oxygens in the hydroxamate moiety of the cocrystallized inhibitor ilomastat, forming the pentacoordinate system. Similarly, the catalytic zinc in stromelysin-1 (MMP-3) is coordinated by three histidine residues, and in 1SLN the two additional coordination sites are occupied by hydroxamate oxygens in the cocrystallized inhibitor. The structural zinc centers in the MMP-1 and Zn-finger enzymes are both tetracoordinate, with the four coordination sites occupied, respectively, by three histidines and one aspartic acid (2CLT), and four cysteine residues (2J6A).

The metal−ligand distances in the metalloproteins are often quite different from those in small gas-phase molecules. For example, the Zn−N distances in LF are 2.25−2.32 Å and those in MMP-3 are 1.78−1.87 Å, as compared to 1.94−2.06 Å in $Zn(NH_3)_n^{2+}$. The Zn−O distances in the crystal structure sites studied here are 2.24, 2.23, and 2.52 Å (LF); 2.08 Å (MMP-1); and 1.75 and 2.87 Å (MMP-3) vs 1.76−1.85 Å in $Zn(OH)_n^{+2-n}$ and $Zn(H_2O)^{2+}$. The Zn−S distances in the Zn-finger structural site are 2.295−2.40 Å vs 2.17 Å in $Zn(SCH_3)^+$. Note that the geometries of the crystal-site compounds are *not* optimized in the gas phase. The Zn−ligand distances were fixed at their values in the crystals, and the biological ligands were represented as follows: imidazoles were replaced with ammonia, cysteines were replaced with SCH_3, and each oxygen of glutamate or aspartate was replaced by an OH. The SCH_3 groups were placed so that the S and C were at the same positions as in cysteine. The atoms on NH_3, SCH_3, and OH were placed at standard distances and default orientations by the GaussView program; the default methyl C−H bond length is 1.07 Å, the default N−H bond length in NH_3 is 1.00 Å, and the default OH distance is 0.96 Å. The default bond angles for

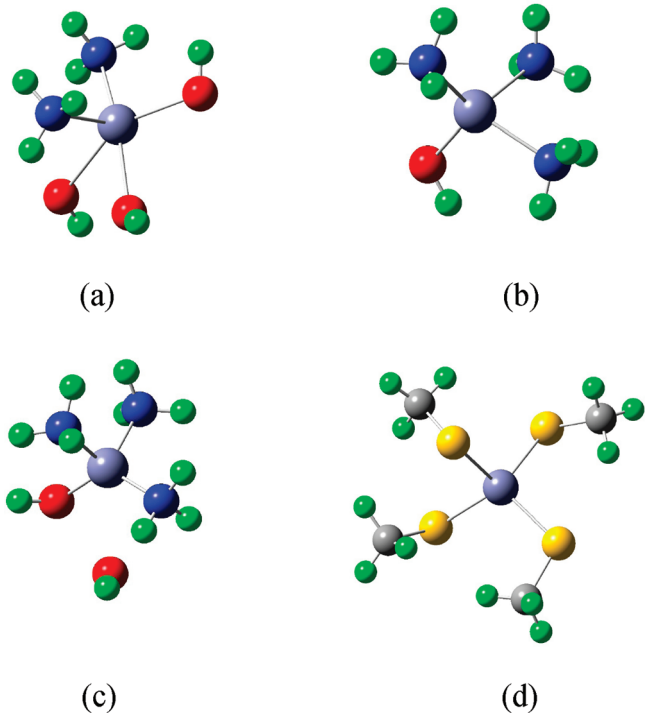


Figure 1. Zn model compounds representing structural and catalytic zinc centers in four relevant metalloproteins: (a) the anthrax toxin lethal factor (LF) (catalytic site),[73] (b) matrix metalloproteinase-1 (MMP-1) (structural site),[74] (c) matrix metalloproteinase-3 (MMP-3) (catalytic site),[75] and (d) a *S. cerevisiae* Zn-finger protein (structural site).[76] Light blue = Zn, red = O, dark blue = N, yellow = S, gray = C, and green = H.

ligand geometries in GaussView are obtained by AM1 optimizations.

Dissociation energies were calculated for each Zn-ligand bond, and dipole moments were calculated for each model compound as follows: first at the crystal structure geometry and then at the Zn-containing product of the dissociation process for breaking each Zn-ligand bond. In addition, we calculated the change in energy when the bond distance (to Zn) of each ligand was stretched by 0.2 Å as compared to its value in the X-ray structure. (The stretch was carried out along the bond axis with the remainder of the molecule held rigid.) The calculations were done using the same DFT and NDDO methods that were tested for small molecules, and the results were compared with CCSD(T)/B2 values for energies and CCSD/B2 values for dipole moments for the LF and MMP-1 crystal structure site models. In addition, the NDDO methods were tested over all four X-ray structures against benchmark values of bond dissociation energies and dipole moments, including M05-2X values for the MMP-3 and Zn-finger crystal structure models.

4. Results and Discussion

4.1. Small Molecules and Clusters. Table 1 shows that the B3LYP functional is the most accurate one for reproducing Zn−ligand bond lengths, particularly Zn−Zn bond distances. The energetic surface of Zn−Zn bonds in the vicinity of the optimized value is very flat, i.e., large changes in bond distance often lead to a very small energy change. The mean unsigned error in Zn−Zn bond length is therefore quite large in comparison with MUEs for other Zn−ligand bonds. Therefore in Table 1 we also show bond-distance MUEs excluding Zn−Zn bonds. If one excludes compounds with Zn−Zn bonds, the MUE in bond distance for B3LYP differs from M05-2X by only 0.002 Å.

Among the NDDO methods, ZnB shows the best perfomance for bond lengths evaluated against the D1 database, but Table 9 shows that its performance is not superior to other methods if compounds containing Zn−Zn bonds are excluded. The ZnB parametrization accurately reproduces the lengths of Zn−S bonds, but not Zn−O bond lengths. The most accurate NDDO method for Zn−ligand bond lengths without Zn−Zn bonds is PM3. PM3(tm) improves upon PM3 for Zn−O and Zn−S bonds, but does not do well for Zn−N bonds. However, AM1 performs particularly well for Zn−N bonds. As was determined in previous work,[89] we found that MNDO/d does not accurately reproduce Zn−S bond distances. The PM6 method is generally not accurate for our test data. Table 10 illustrates that the broadening of our test set to larger systems by inclusion of M05-2X bond lengths does not alter these conclusions.

The best NDDO methods for Zn−ligand bond angles in our training set are PM3 and AM1, while the least accurate method is PM6 (Table 11). For the enlarged benchmark database for bond angles, the PM3 method no longer exhibits the best performance, and the best methods against the D2 database benchmarks are AM1 and ZnB (which is also shown in Table 11).

The best density functional for dipole moments relative to the CCSD benchmarks is M05-2X (Table 5), followed closely by M06-2X, while the best NDDO method for dipole moments is PM3 (Table 12). Adding CCSD//M05-2X benchmarks (D2 database) does not change the leadership of M05-2X (Table 12). DFT methods with a lower percentage X of Hartree−Fock exchange, including B3LYP and M06, are less accurate for dipole moments than M06-2X and M05-2X (X = 56% and 54%, respectively). PM3 exhibits

Table 9. Mean Unsigned Errors (MUEs) in Zn−Ligand Bond Distances (in Å) for Various NDDO Methods Relative to D1 Benchmark Database Values[a]

method	Zn−H	Zn−C	Zn−N	Zn−O	Zn−S	Zn−halogen	Zn−Zn	all compds	excluding Zn−Zn bonds
ZnB	0.12	0.03	0.04	0.11	0.08	0.09	0.57	0.14	0.08
MNDO/d	0.12	0.08	0.04	0.04	0.15	0.02	0.93	0.17	0.07
PM3	0.02	0.04	0.04	0.04	0.16	0.02	1.18	0.18	0.05
PM3(tm)	0.04	0.04	0.08	0.03	0.11	0.03	1.21	0.18	0.05
AM1	0.06	0.07	0.02	0.09	0.07	0.03	1.11	0.18	0.06
PM6	0.13	0.08	0.07	0.15	0.10	0.15	1.54	0.28	0.12

[a] MUE is based on 26 CCSD(T) values in Table 1.

Table 10. Mean Unsigned Errors (MUEs) in Zn-Ligand Bond Distances (in Å) for Various NDDO Methods Relative to Augmented Benchmark Database Values (database D2)[a]

method	Zn−H	Zn−C	Zn−N	Zn−O	Zn−S	Zn−halogen	Zn−Zn	all compds	excluding Zn−Zn bonds
ZnB	0.12	0.03	0.07	0.09	0.09	0.09	0.57	0.11	0.08
PM3	0.02	0.04	0.06	0.04	0.09	0.02	0.93	0.13	0.05
MNDO/d	0.12	0.08	0.06	0.04	0.14	0.03	1.11	0.13	0.07
PM3(tm)	0.04	0.04	0.12	0.03	0.09	0.02	1.18	0.14	0.06
AM1	0.06	0.07	0.03	0.11	0.05	0.03	1.21	0.14	0.07
PM6	0.13	0.10	0.11	0.14	0.06	0.15	1.54	0.21	0.11

[a] MUE is based on 43 CCSD(T) and M05-2X values in Tables 1 and 2.

Table 11. Mean Unsigned Errors (MUEs) in Zn−Ligand Bond Angles (in deg) for Various NDDO Methods, Relative to Databases D1 and D2

method	D1[a]	D2[b]
AM1	3.0	4.3
ZnB	3.6	4.8
MNDO/d	8.6	6.2
PM3	2.4	6.8
PM3(tm)	6.9	8.4
PM6	10.4	10.1

[a] MUE for database D1 is relative to nine CCSD(T) values in Table 3. [b] MUE for augmented database D2 is relative to 38 CCSD(T) and M05-2X values in Tables 3 and 4.

Table 12. Mean Unsigned Errors (MUEs) in Dipole Moments (D) for Various DFT and NDDO Methods, Relative to Databases D1 and D2

method	D1[a]	D2[b]
M05-2X	0.14	0.14
M06-2X	0.16	0.15
M06	0.36	0.36
B3LYP	0.39	0.37
PM3	0.80	0.86
AM1	0.89	0.91
MNDO/d	0.94	1.07
ZnB	1.16	1.17
PM3(tm)	1.47	1.52
PM6	1.94	2.02

[a] MUE for database D1 is relative to 20 CCSD values in Table 5. [b] MUE for database D2 is relative to 22 CCSD and CCSD//M05-2X values in Tables 5 and 6.

Table 13. Mean Unsigned Errors (MUEs) in Bond Dissociation Energies (kcal/mol) for Various DFT and NDDO Methods, Relative to Databases D1 and D2

method	D1[a]	D2[b]
M05-2X	3.1	3.7
M06-2X	3.4	3.8
M06	5.5	5.2
B3LYP	6.0	6.0
MNDO/d	14.2	13.8
ZnB	18.1	18.3
PM3	21.5	20.5
AM1	29.6	27.8
PM6	26.8	28.9
PM3(tm)	71.6	65.1

[a] MUE for database D1 is relative to 29 CCSD(T) values in Table 7. [b] MUE for database D2 is relative to 37 CCSD(T) and CCSD(T)//M05-2X values in Tables 7 and 8.

the smallest MUE among the semiempirical methods for the large database (D2) of dipole moments. In our previous work, PM3(tm) was not evaluated for dipole moments, because it was not parametrized for these, but our present calculations show that its performance for dipole moments is indeed quite poor. As the dipole moment may be seen as the leading indicator of charge distribution, one must view with caution a method that does not yield reasonable dipole moments.

Table 14. Balanced Mean Unsigned Errors (BMUEs) for Zn-Containing Molecules and Clusters, Based on Comparing DFT and NDDO Bond Distances, Bond Dissociation Energies, and Dipole Moments to CCSD(T) Bond Distances, CCSD(T) Bond Dissociation Energies, and CCSD Dipole Moments Values in the D1 Database

method	molecules[a]	molecules and clusters[b]
M05-2X	0.21	0.21
M06-2X	0.29	0.29
B3LYP	0.35	0.32
M06	0.35	0.41
MNDO/d	1.08	1.05
ZnB	1.30	1.11
PM3	1.01	1.13
AM1	1.26	1.30
PM6	2.03	1.93
PM3(tm)	2.13	2.25

[a] BMUE calculated excluding Zn−Zn bond lengths. [b] BMUE calculated including Zn−Zn bond lengths.

Table 15. Balanced Mean Unsigned Errors (BMUEs) in NDDO Bond Distances, Bond Angles, Bond Dissociation Energies, and Dipole Moments for Zn-Containing Molecules and Clusters, Validated against the Enlarged Benchmark Database (D2)

	BMUE	
method	excluding Zn−Zn[a]	including Zn−Zn[b]
ZnB	0.84	0.76
MNDO/d	0.80	0.79
PM3	0.76	0.82
AM1	0.81	0.82
PM6	1.42	1.40
PM3(tm)	1.37	1.41

[a] BMUE calculated excluding Zn−Zn bond lengths. [b] BMUE calculated including Zn−Zn bond lengths.

M05-2X is also the top DFT method for bond dissociation energies (Table 7), with M06-2X in second place. MNDO/d predicts the most accurate dissociation energies among NDDO methods tested (Table 13). The new ZnB parametrization of PM3 for Zn also shows very good performance for dissociation energies. These results differ from the conclusion drawn in ref 112, where PM6 is reported as the most accurate method for dissociation energies. Testing bond dissociation energies against the augmented benchmark database D2, which includes CCSD(T)//M05-2X values (Table 13), does not change the ranking of DFT methods,

Table 16. Balanced Mean Unsigned Errors (BMUEs) in DFT and NDDO Dissociation Energies and Dipole Moments for Zn-Containing Molecules and Clusters, Evaluated against the Enlarged Benchmark Database (D2)

method	BMUE
M05-2X	0.18
M06-2X	0.18
M06	0.34
B3LYP	0.37
MNDO/d	0.98
PM3	1.03
ZnB	1.16
AM1	1.25
PM6	1.92
PM3(tm)	2.57

[a] Based on 81 geometric data, 36 energetic data, and 22 dipole moments.

but the rank of PM6 is improved, and its MUE becomes smaller than the MUE for AM1.

The balanced mean unsigned error (BMUE) is a unitless quantity that normalizes MUEs for each parameter against the average error over all methods for that parameter. Thus, BMUE serves as a criterion to evaluate the overall performance of each technique. For the sake of simplicity, we first restricted BMUE to incorporate bond distances, dissociation energies, and dipole moments for database D1. This initial BMUE is defined as

$$\mathrm{BMUE} = \{[\mathrm{MUE\ (in\ \AA)}/\mathrm{AMUE\ (in\ \AA)}] + [\mathrm{MUE\ (in\ D)}/\mathrm{AMUE\ (in\ D)}] + [\mathrm{MUE\ (inkcal/mol)}/\mathrm{AMUE\ (in\ kcal/mol)}]\}/3 \quad (1)$$

where AMUE is the average mean unsigned error, i.e., the mean of all MUEs for bond distances (in Å), dipole moments (in D), or bond dissociation energies (in kcal/mol). Table 14 lists BMUEs for all methods tested in this paper calculated relative to CCSD(T) bond distances and dissociation energies and relative to CCSD dipole moments (that is, over the D1 database). M05-2X is found to be the best density functional, while MNDO/d is the best NDDO method (mainly due to highly accurate bond dissociation energies in comparison with other NDDO methods). Among the NDDO methods, ZnB also demonstrates very good performance. M06-2X is found to be in second place overall. Excluding Zn−Zn bond distances from the BMUE calculations does not change the ranking of density functionals; however, without Zn−Zn bond distances, MNDO/d and ZnB rank below PM3.

For database D2, we incorporated bond angles into eq 1 by adding a fourth term, MUE (in deg)/AMUE (in deg), to the average. BMUEs calculated relative to the D2 augmented benchmark database for Zn−ligand bond distances and angles, dipole moments, and dissociation energies are listed in Table 15. ZnB proves superior here due to its accurate rendering of Zn−Zn bonds and dissociation energies; however, removing the Zn−Zn bond distances from the test set raises the PM3 method to first place, as it ranks best for bond distances (without Zn−Zn bonds) and for dipole moments. Despite its good perfomance for bond distances, PM3(tm) fails to give accurate dissociation energies or dipole moments; therefore, its overall rank is very low. Table 16 lists BMUEs of all tested methods against the D2 database

Table 17. Balanced Mean Unsigned Errors (BMUEs) for Two Coordination Compounds Based on Metalloenzyme Crystal Structures (LF and MMP-1) Determined by Comparing DFT and NDDO Bond Dissociation Energies, Bond-Stretch Energies, Dipole Moments, and Fragment Dipole Moments to CCSD(T) Dissociation Energies and CCSD Dipole Moments[a]

	MUE		
method	dipole moment (D)	dissociation energy (kcal/mol)	BMUE
M06-2X	0.04	3.2	0.32
M05-2X	0.09	2.8	0.33
M06	0.38	3.1	0.72
PM3	0.33	4.7	0.78
MNDO/d	0.23	6.6	0.82
ZnB	0.50	5.0	1.02
B3LYP	0.42	6.5	1.04
AM1	0.47	6.0	1.15
PM6	0.85	5.5	1.48
PM3(tm)	0.79	17.1	2.36

[a] Based on 18 energetic data and 11 dipole moments.

Table 18. Mean Unsigned Errors (MUEs) and Balanced Mean Unsigned Errors (BMUEs) for Four Coordination Compounds Representing Metalloenzyme Crystal Structure Zn Centers (LF, MMP-1, MMP-3, and Zn-finger), Incorporating Bond Dissociation Energies, Bond-Stretch Energies, Dipole Moments, and Fragment Dipole Moments[a]

	MUE		
method	dipole moment (D)	dissociation energy (kcal/mol)	BMUE
PM3	0.44	5.7	0.60
ZnB	0.81	6.3	0.85
AM1	0.90	6.7	0.92
PM6	0.99	7.1	1.00
MNDO/d	0.97	9.3	1.12
PM3(tm)	0.99	15.5	1.50

[a] Based on 36 energetic data and 22 dipole moments.

when BMUE is restricted to dissociation energies and dipole moments; once again, the ranking of density functionals does not change, but now MNDO/d ranks above PM3 due to its excellent performance for bond dissociation energies.

4.2. Metalloprotein Zn Center Models. The MUEs for dissociation energies and dipole moments and corresponding BMUEs against CCSD(T) values of dissociation energies and bond-stretch energies and CCSD values of dipole moments for the Zn-center models obtained from the LF and MMP-1 X-ray structures are given in Table 17. These tests are consistent with the results of section 3.1 in that the best density functionals are M05-2X and M06-2X, but B3LYP shows surprisingly poor performance for the large crystal-site coordination compounds; its rank is lower than the ranks of three NDDO methods (PM3, MNDO/d, and ZnB). The poor performance of B3LYP in these cases casts doubts on methods such as SCC-DFTB[131] that were parametrized[132] against B3LYP for Zn-containing compounds.

Including M05-2X benchmark values of dissociation energies, bond-stretch energies, and dipole moments, we calculated MUEs and BMUEs for all four crystal-structure model compounds (Table 18). These results support our

conclusion that the most suitable NDDO method for Zn compounds is PM3.

4.3. Other Transition Metals. When M05-2X[92] and M06-2X[99] were developed, it was specifically indicated that they were not designed to treat transition metals because the accurate treatment of the multireference character[133] of transition metals usually requires 70−100% local exchange and no more than 30% Hartree−Fock exchange. Nevertheless, in our previous paper[89] we tested these functionals and found that M05-2X gives the best performance for Zn chemistry. Perhaps this should not be too surprising, because the closed shell ($3d^{10}4s^2$) of Zn means that it has little multireference character in many of its complexes. Apparently Zn acts more like a main-group metal than a transition metal in its terms of its requirements for accurate electronic structural treatment. Recent tests and discussions of tests of density functionals for complexes of other transition metals may be found elsewhere.[92,99,100,134−144]

5. Summary and Concluding Remarks

We have presented databases of CCSD(T)/B2 geometric parameters and bond dissociation energies and CCSD/B2 dipole moments for a set of Zn model compounds. This database, D1, is supplemented by M05-2X/B2 geometries, CCSD(T)/B2 single-point dissociation energies, and CCSD single-point dipole moments calculated at M05-2X geometries, forming a larger database called D2. Four density functionals and six semiempirical parametrizations of the NDDO method were tested relative to all benchmark values.

As expected, the density functionals we tested prove much more accurate than the semiempirical NDDO methods for the small-molecule Zn compounds. For all calculated features of small compounds, the BMUE for the worst density functional is at least a factor of 2 smaller than the BMUE for the best NDDO method. This also holds true for the metalloprotein site models with exception of B3LYP, which does not perform as well as PM3, MNDO/d, or ZnB for these structures.

Our conclusions agree with those of ref 89 in that M05-2X is the most suitable density functional method for accurate calculations of geometries, dipole moments, and energetics of Zn compounds. This conclusion is further supported by the fact that the database used for testing in the present study is now larger, more diverse, and more broadly representative of the larger structures that occur in important Zn environments such as clusters and metalloproteins.

PM3 and MNDO/d are found to be, on average, the best NDDO methods for describing Zn centers. The recently proposed[84] ZnB parametrization of PM3 is an improvement upon PM3 for Zn−Zn bond distances and energies, but in general does not improve upon that method for other geometrical parameters and dipole moments.

Acknowledgment. This work was supported in part by the University of Minnesota Department of Medicinal Chemistry, by the University of Minnesota Academic Health Center, and by the National Science Foundation under Grant No. CHE07-04974.

Supporting Information Available: CCSD(T) and M05-2X Cartesian geometries of all Zn−ligand compounds studied in this paper; DFT and NDDO Zn−ligand bond distances, bond angles, dipole moments and bond dissociation energies; X-ray structural coordinates for the Zn biocenters studied in this paper; bond dissociation energies, bond-stretch energies, and dipole moments of coordination compounds based on X-ray structures and their dissociation products. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Vargas-Tah, A. A.; Garcia, R. C.; Archila, R. F. P.; Solis, J. R.; López, A. J. C. *Catal. Today* **2005**, *107*, 713.

(2) Di Serio, M.; Tesser, R.; Dimiccoli, M.; Cammarota, F.; Nastasi, M.; Santasesaria, E. *J. Mol. Catal. A* **2005**, *239*, 111.

(3) De Cola, P. L.; Gläser, R.; Weitkamp, J. *Appl. Catal., A* **2006**, *306*, 85.

(4) Yin, C.; Zhao, R.; Zhao, H.; Xu, Y.; Liu, C. *Pet. Sci. Technol.* **2007**, *25*, 491.

(5) Baruwati, B.; Kumar, D. K.; Manorama, S. V. *Sensors Actuators, B* **2006**, *119*, 676.

(6) Srivastava, A.; Rashmi; Jain, K. *Mater. Chem. Phys.* **2007**, *105*, 385.

(7) Shinde, V. R.; Gujar, T. P.; Lokhande, C. D.; Mane, R. S.; Han, S.-H. *Mater. Sci. Eng., B* **2007**, *137*, 119.

(8) Shinde, V. R.; Gujar, T. P.; Lokhande, C. D.; Mane, R. S.; Han, S, H. *Sensors Actuators, B* **2007**, *123*, 882.

(9) Kapse, V. D.; Ghosh, S. A.; Raghuwanshi, F. C.; Kapse, S. D. *Mater. Chem. Phys* **2009**, *113*, 638.

(10) Mishra, D.; Srivastava, A.; Srivastava, A.; Shukla, R. K. *Appl. Surf. Sci.* **2008**, *255*, 2947.

(11) Callan, J. F.; de Silva, A. P.; Margi, D. C. *Tetrahedron* **2005**, *61*, 8551.

(12) Kikuchi, K.; Komatsu, K.; Nagano, T. *Curr. Opin. Chem. Biol.* **2004**, *8*, 182.

(13) Wang, J. *Analyst* **2005**, *130*, 421.

(14) Dai, Z.; Canary, J. W. *New J. Chem.* **2007**, *31*, 1708.

(15) Leroux, J. C.; Allémann, E.; De Jaeghere, F.; Doelker, E.; Gurny, R. *J. Controlled Release* **1996**, *39*, 339.

(16) Huang, M. H.; Mao, S.; Feick, H.; Yan, H.; Wu, Y.; Kind, H.; Weber, E.; Russo, R.; Yang, P. *Science* **2001**, *292*, 1897.

(17) Tani, T.; Mädler, L.; Pratsinis, S. E. J. *Nanopart. Res.* **2002**, *4*, 337.

(18) Greene, L. E.; Law, M.; Goldberger, J.; Kim, F.; Johnson, J. C.; Zhang, Y.; Saykally, R. J.; Yang, P. *Angew. Chem., Int. Ed.* **2003**, *42*, 3031.

(19) Vayssieres, L. *Adv. Mater.* **2003**, *15*, 464.

(20) Liu, B.; Zeng, H. C. *J. Am. Chem. Soc.* **2003**, *125*, 4430.

(21) Kong, X. Y.; Wang, Z. L. *Nano Lett.* **2003**, *3*, 1625.

(22) Wang, R. M.; Xing, Y. J.; Xu, J.; Yu, D. P. *New J. Phys.* **2003**, *5*, 115.

(23) Larson, D. R.; Zipfel, W. R.; Williams, R. M.; Clark, S. W.; Bruchez, M. P.; Wise, F. W.; Webb, W. W. *Science* **2003**, *300*, 1434.

(24) Beek, W. J.; Wienk, M. M.; Janssen, R. A. J. *Adv. Mater.* **2003**, *16*, 1009.

(25) Su, C.-Y.; Goforth, A. M.; Smith, M. D.; Pellechia, P. J.; zur Loye, H.-C. *J. Am. Chem. Soc.* **2004**, *126*, 3576.

(26) Kong, X. Y.; Ding, Y.; Yang, R.; Wang, Z. L. *Science* **2004**, *303*, 1348.

(27) Li, Q.; Kumar, V.; Li, Y.; Zhang, H.; Marks, T. J.; Chang, R. P. H. *Chem. Mater.* **2005**, *17*, 1001.

(28) Gao, P. X.; Ding, Y.; Mai, W.; Hughes, W. L.; Lao, C. S.; Wang, Z. L. *Science* **2005**, *309*, 1700.

(29) Lao, C. S.; Liu, J.; Gao, P. X.; Zhang, L.; Davidovic, D.; Tummala, R.; Wang, Z. L. *Nano Lett.* **2006**, *6*, 263.

(30) Fu, Z.; Dong, W.; Yang, B.; Wang, Z.; Yang, Y.; Yan, H.; Zhang, S.; Zuo, J.; Ma, M.; Liu, X. *Solid State Commun.* **2006**, *138*, 179.

(31) Pradhan, N.; Battaglia, D. M.; Liu, Y.; Peng, X. *Nano Lett.* **2007**, *7*, 312.

(32) Soci, C.; Zhang, A.; Xiang, B.; Dayeh, S. A.; Aplin, D. P. R.; Park, J.; Bao, X. Y.; Lo, Y. H.; Wang, D. *Nano Lett.* **2007**, *7*, 1003.

(33) Leschkies, K. S.; Divakar, R.; Basu, J.; Enache-Pommer, E.; Boercker, J. E.; Carter, C. B.; Kortshagen, U. R.; Norris, D. J.; Aydil, E. S. *Nano Lett.* **2007**, *7*, 1793.

(34) Kee, C. S.; Ko, D. K.; Lee, J. *Solid State Commun.* **2007**, *142*, 195.

(35) McCall, K. A.; Huang, C.-C.; Fierke, C. A. *J. Nutr.* **2000**, *130*, 1437S.

(36) Amin, E. A.; Welsh, W. J. *J. Med. Chem.* **2001**, *44*, 3849.

(37) Kilshtain-Vardi, A.; Shoham, G.; Goldblum, A. *Int. J. Quantum Chem.* **2002**, *88*, 87.

(38) Bräuer, M.; Pérez-Lustres, J. L.; Weston, J.; Anders, E. *Inorg. Chem.* **2002**, *41*, 1454.

(39) Hou, T.; Zhang, W.; Xu, X. *J. Computer-Aided Mol. Des.* **2002**, *16*, 27.

(40) Hasegawa, K.; Ono, T.; Noguchi, T. *J. Phys. Chem. A* **2002**, *106*, 3377.

(41) Krauss, M.; Olsen, L.; Antony, J.; Hemmingsen, L. *J. Phys. Chem. B* **2002**, *106*, 9446.

(42) Dudev, T.; Lim, C. *Chem. Rev.* **2003**, *103*, 773.

(43) Dudev, T.; Lin, Y.; Dudev, M.; Lim, C. *J. Am. Chem. Soc.* **2003**, *125*, 3168.

(44) Bergquist, C.; Fillebeen, T.; Morlok, M. M.; Parkin, J. *J. Am. Chem. Soc.* **2003**, *125*, 6189.

(45) Babu, C. S.; Dudev, T.; Casareno, R.; Cowan, J. A.; Lim, C. *J. Am. Chem. Soc.* **2003**, *125*, 9318.

(46) Hu, X.; Shelver, W. H. *J. Mol. Graphics Modelling* **2003**, *22*, 115.

(47) Krauss, M.; Gresh, N.; Antony, J. *J. Phys. Chem. B* **2003**, *107*, 1215.

(48) Gervasio, F. L.; Schettino, V.; Mangani, S.; Krack, M.; Carloni, P.; Parrinello, M. *J. Phys. Chem. B* **2003**, *107*, 6886.

(49) Remko, M.; Garaj, V. *Mol. Phys.* **2003**, *101*, 2357.

(50) Gleeson, M. P.; Burton, N. A.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4272.

(51) Parkin, G. *Chem. Rev.* **2004**, *104*, 699.

(52) Brown, D. A.; Cuffe, L. P.; Fitzpatrick, N. J.; Ryan, A. T. *Inorg. Chem.* **2004**, *43*, 297.

(53) Asthagiri, D.; Pratt, L. R.; Paulaitis, M. E.; Rempe, S. B. *J. Am. Chem. Soc.* **2004**, *126*, 1285.

(54) Xu, Q.; Guo, H. *J. Phys. Chem. B* **2004**, *108*, 2477.

(55) Linder, D. P.; Rodgers, K. R. *J. Phys. Chem. B* **2004**, *108*, 13839.

(56) Ambroggio, X. A.; Rees, D. C.; Deshaies, R. J. *PLoS Biol.* **2004**, *2*, 0113.

(57) Fleischer, H. *Coord. Chem. Rev.* **2005**, *249*, 799.

(58) Guo, H.; Rao, N.; Xu, Q.; Guo, H. *J. Am. Chem. Soc.* **2005**, *127*, 3191.

(59) Dudev, T.; Chang, L.-Y.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4091.

(60) Marino, T.; Russo, N.; Toscano, M. *J. Am. Chem. Soc.* **2005**, *127*, 4242.

(61) Sakharov, D. V.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921.

(62) Lin, Y.-L.; Lee, Y.-M.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 11336.

(63) Bozym, R. A.; Thompson, R. B.; Stoddard, A. K.; Fierke, C. A. *ACS Chem. Biol.* **2006**, *1*, 103.

(64) Petros, A. K.; Reddi, A. R.; Kennedy, M. L.; Hyslop, A. G.; Gibney, B. R. *Inorg. Chem.* **2006**, *45*, 9941.

(65) Estiu, G.; Suárez, D.; Merz, K. M., Jr. *J. Comput. Chem.* **2006**, *27*, 1240.

(66) Furukawa, Y.; Fu, R.; Deng, H.-X.; Siddique, T.; O'Halloran, T. V. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7148.

(67) Reddi, A. R.; Gibney, B. R. *Biochemistry* **2007**, *46*, 3745.

(68) Gupta, S. P. *Chem. Rev.* **2007**, *107*, 3042.

(69) Riccardi, D.; Cui, Q. *J. Phys. Chem. A* **2007**, *111*, 5703.

(70) Poznański, J.; Najda, A.; Bretner, M.; Shugar, D. *J. Phys. Chem. A* **2007**, *111*, 6501.

(71) Lee, Y.-M.; Lim, C. *J. Mol. Biol.* **2008**, *379*, 545.

(72) Koutmos, M.; Pejchal, R.; Bomer, T. M.; Matthews, R. G.; Smith, J. L.; Ludwig, M. L. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3286.

(73) Turk, B. E.; Wong, T. Y.; Schwarzenbacher, R.; Jarrell, E. T.; Leppla, S. H.; Collier, R. J.; Liddington, R. C.; Cantley, L. C. *Nat. Struct. Mol. Biol.* **2004**, *11*, 60.

(74) Iyer, S.; Visse, R.; Nagase, H.; Acharya, K. R. *J. Mol. Biol.* **2006**, *362*, 78.

(75) Becker, J. W.; Marcy, A. I.; Rokosz, L. L.; Axel, M. G.; Burbaum, J. J.; Fitzgerald, P. M.; Cameron, P. M.; Esser, C. K.; Hagmann, W. K.; Hermes, J. D.; Springer, J. P. *Protein Sci.* **1995**, *4*, 1966.

(76) Heurgue-Hamard, V.; Graille, M.; Scrima, N.; Ulryck, N. *J. Biol. Chem.* **2006**, *281*, 36140.

(77) Rude, U. *Proteins: Struct. Funct. Genet.* **1995**, *21*, 40.

(78) Kafafi, S. A.; Krauss, M. *Int. J. Quantum Chem.* **1999**, *75*, 289.

(79) Bräuer, M.; Kunert, M.; Dinjus, E.; Kluβmann, M.; Döring, M.; Görls, H.; Anders, E. *J. Mol. Struct. (THEOCHEM)* **2000**, *505*, 289.

(80) Yazal, E. J.; Pang, Y. P. *J. Mol. Struct. (THEOCHEM)* **2001**, *545*, 271.

(81) Matxain, J. M.; Fowler, J. E.; Ugalde, J. M. *Phys. Rev. A* **2000**, *61*, 053201.

(82) Matxain, J. M.; Fowler, J. E.; Ugalde, J. M. *Phys. Rev. A* **2000**, *62*, 053201.

(83) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565.

(84) Brothers, E. N.; Suarez, D.; Deerfield, D. W., II.; Merz, K. M., Jr. *J. Comput. Chem.* **2004**, *25*, 1677.

(85) Tachikawa, H.; Iokibe, K.; Azumi, K.; Kawabata, H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3978.

(86) Iokibe, K.; Tachikawa, H.; Azumi, K. *J. Phys. B: At. Mol. Opt. Phys.* **2007**, *40*, 427.

(87) Drew, M. G. B.; Parui, D.; De, S.; Chowdhury, S.; Dipankar, D. *New J. Chem.* **2007**, *31*, 1763.

(88) Jain, T.; Jayaram, B. *Proteins: Struct. Funct. Bioinf.* **2007**, *67*, 1167.

(89) Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 75.

(90) Frison, G.; Ohanessian, G. *J. Comput. Chem.* **2008**, *29*, 416.

(91) Jiang, L.; Xu, Q. *J. Chem. Phys.* **2008**, *128*, 124317.

(92) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

(93) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.

(94) Bartlett, R. J. *J. Phys. Chem.* **1989**, *93*, 1697.

(95) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(96) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(97) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(98) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(99) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *120*, 215.

(100) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(101) Dewar, M. J. S.; Merz, K. M., Jr. *Organometallics* **1988**, *7*, 522.

(102) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(103) Dewar, M. J. S.; Jie, C.; Yu, G. *Tetrahedron* **1993**, *23*, 5003.

(104) Holder, A. J.; Dennington, R. D.; Jie, C.; Yu, G. *Tetrahedron* **1994**, *50*, 627.

(105) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

(106) Hehre, W. J.; Yu, J.; Adei, E. *Abstracts of Papers, 212th ACS National Meeting, Orlando, FL, Aug. 25−29, 1996*; American Chemical Society: Washington, DC, 1996; Abstract COMP 092.

(107) Hehre, W. J. *A Guide to Molecular Mechanics and Quantum Chemical Calculations*; Wavefunction, Inc.: Irvine, CA, 2003; p 48.

(108) Cundari, T. R.; Deng, J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 376.

(109) Bosque, R.; Maseras, F. *J. Comput. Chem.* **2000**, *21*, 562.

(110) Thiel, W.; Voityuk, A. *Theor. Chim. Acta* **1992**, *81*, 391.

(111) Thiel, W.; Voityuk, A. *J. Phys. Chem.* **1996**, *100*, 616.

(112) Stewart, J. J. P. *J. Mol. Modeling* **2007**, *13*, 1173.

(113) Anh, N. T.; Frison, G.; Solladié-Cavallo, A.; Metzner, P. *Tetrahedron* **1998**, *54*, 12841.

(114) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.

(115) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907.

(116) Wachters, A. J. H. *J. Chem. Phys.* **1970**, *52*, 1033.

(117) Hay, P. J. *J. Chem. Phys.* **1977**, *68*, 4377.

(118) Raghavachari, K.; Trucks, G. W. *J. Chem. Phys.* **1989**, *91*, 1062.

(119) 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, N.; 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, 2004.

(120) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866.

(121) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem A* **2003**, *107*, 1384.

(122) Schuchardt, K. L.; Didler, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045.

(123) Zhao, Y.; Truhlar, D. G. *MN-GFM: Minnesota Gaussian Functional Module—Version 3.0*; University of Minnesota: Minneapolis, MN, 2006.http://comp.chem.umn.edu/mn-gfm/, accessed 6/5/2008.

(124) Dewar, M. J. S.; Stewart, J. J. P.http://www.openmopac.net/index.html, accessed 2/11/2008.

(125) Stewart, J. J. P.; Zheng, J.; Rossi, I.; Hu, W.-P.; Lynch, G. C.; Liu, Y.-P.; Chuang, Y. Y.; Pu, J.; Li, J.; Cramer, C. J.; Fast, P. L.; Sorkin, A.; Truhlar, D. G. *MOPAC 5.012mn*; University of Minnesota, 2008; http://comp.chem.umn.edu/mopac/, accessed 3/9/2009.

(126) *Spartan'04*; Wavefunction, Inc., Irvine, CA; http://www.wavefun.com/products/spartan.html, accessed 3/14/2006.

(127) Kinal, A.; Piecuch, P. *J. Phys. Chem. A* **2007**, *111*, 734.

(128) Munzarova, M.; Kaupp, M. *J. Phys. Chem. A* **1999**, *103*, 9966.

(129) Li, X.; Paldus, J. *J. Chem. Phys.* **2007**, *126*, 234303.

(130) Mayor-López, M. J.; Lüthi, H. P.; Koch, H.; Morgantini, P. Y.; Weber, J. *J. Chem. Phys.* **2000**, *113*, 8009.

(131) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.

(132) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565.

(133) Truhlar, D. G. *J. Comput. Chem.* **2007**, *28*, 73.

(134) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127.

(135) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(136) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103.

(137) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *124*, 224105.

(138) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(139) Harvey, J. N. *Annu. Rep. Prog. Chem., Sect. C,* **2006**, *102*, 203.

(140) Sears, J. S.; Sherrill, C. D. *J. Phys. Chem. A* **2008**, *112*, 3466.

(141) Sears, J. S.; Sherrill, C. D. *J. Phys. Chem. A* **2008**, *112*, 6741.

(142) Ghosh, A.; Gonzalez, E.; Tangen, E.; Roos, B. O. *J. Phys. Chem. A* **2008**, *112*, 12792.

(143) Rinaldo, D.; Tian, L.; Harvey, J. N.; Friesner, R. A. *J. Chem. Phys.* **2008**, *129*, 164108.

(144) Bühl, M.; Reimann, C.; Pantazis, D.; Bredow, T.; Neese, F. *J. Chem. Theory Comput.* **2008**, *4*, 1449.

CT900038M

JCTC Journal of Chemical Theory and Computation

Theoretical Study of Be_N Linear Chains: Optimized Geometries and Harmonic Frequencies

Antonio Monari* and Gian Luigi Bendazzoli

Dipartimento di Chimica Fisica e Inorganica, Università di Bologna Viale Risorgimento 4, I-40136 Bologna, Italy

Stefano Evangelisti

Laboratoire de Chimie et Physique Quantiques, Université de Toulouse et CNRS, 118, Route de Narbonne, F-31062 Toulouse Cedex, France

Received February 3, 2009

Abstract: The electronic structure of linear beryllium chains has been theoretically studied at an *ab initio* level. By using a CAS-SCF approach, geometries have been optimized and harmonic frequencies computed. The optimized geometries present almost equal bond lengths, while all the harmonic frequencies are real. This fact indicates the presence of a local minimum, at this level of theory, having a linear geometry. The energy splitting between the singlet ground state, $^1\Sigma_g$, and the quasi-degenerated excited triplet, $^3\Sigma_u$, has been computed at CAS-SCF and MR-CI level. It was found that the splitting goes exponentially to zero as the number of atoms in the chain is increased.

1. Introduction

The interest on linear surface-deposited atomic structures has grown enormously in the last years, particularly since it became possible to synthesize one-dimensional electron systems at stepped surfaces.[1−5] A common feature of these systems is the presence of edge orbitals corresponding to the "surface states" (in a one-electron formalism) introduced by Tamm and Shockley long ago.[6] Edge effects, for instance, have been predicted in graphene nanoribbons,[7,8] and this can produce a large variety of intersting electric and magnetic behaviors.[9] In one-dimensional systems, these states are also called "end states" and have been detected, for instance, in gold atomic chains deposited on silicon surfaces.[5]

In the case of the phenomenological Hückel Hamiltonian, it has been shown that edge orbitals can be present, under particular conditions, at the extremities of a bond-alternating linear polyene.[10] These edge orbitals, that have quasi-degenerate energies, present exponentially decaying amplitudes in the inner atoms of the chain. In recent works, a similar behavior has been theoretically predicted for the case of linear beryllium chains treated at an *ab initio* level.[11,12] In particular, it has been shown that, for internuclear distances close to the equilibrium structure, two partially filled edge orbitals localize at the chain ends. This gives rise to two low-lying quasi-degenered states, a singlet ground state, $^1\Sigma_g$, and a triplet, $^3\Sigma_u$. Their energy splitting goes exponentially to zero as the number of atoms in the chain is increased.

These calculations were performed by assuming equally spaced linear arrangements for the chain geometries. Therefore, it is interesting to perform separate optimization of all the bond distances, in order to see if the system remains stable under these conditions. Moreover, if this is the case, one must characterize the nature of the stationary point obtained through geometry optimization. If all the second derivatives are positive, i.e., the harmonic frequecies are real (and positive), the stationary point is a local minimum on the Potential Energy Surface (PES) of the system. In this case, it is in principle possible to produce this molecular species, although its stability is related to dynamical and thermodynamical aspects. For these reasons, in the present work we present a CAS-SCF investigation of the geometrical structure of these systems by relaxing the equal-bond constrain. It turns out that the optimized geometries are very close to the constrained-minimum ones. Moreover, all the

* Corresponding author e-mail: amonari@ms.fci.unibo.it.

10.1021/ct9001027 CCC: $40.75
Published on Web 03/30/2009

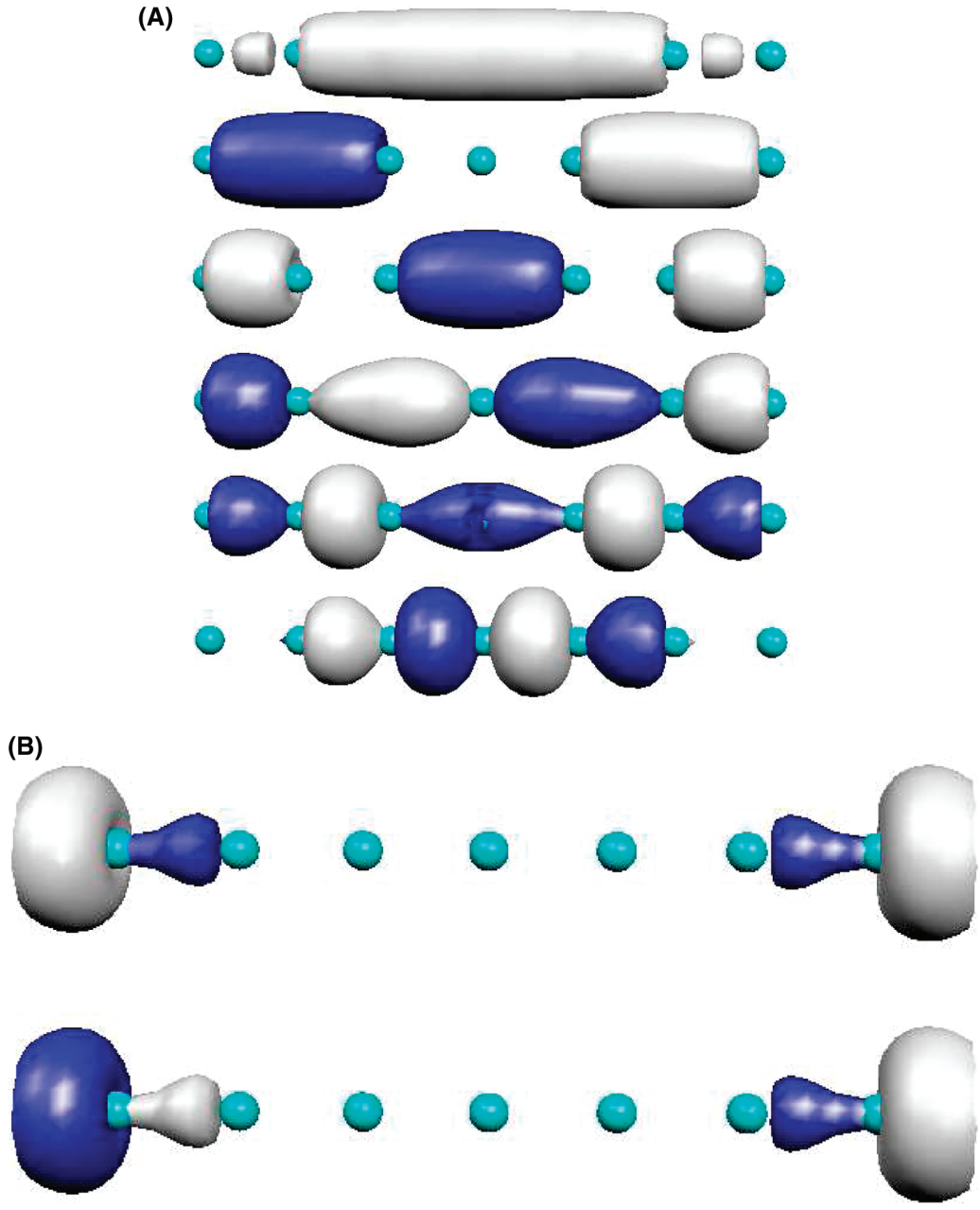


Figure 1. Valence CAS-SCF oritals for the Be_7 chain: occupied orbitals (a) and active orbitals (b).

harmonic frequencies of these systems are real and positive, which means that, at this level of description, the optimized linear structures are local minima on the Potential-Energy Surfaces (PES) of the Be_N chains.

This article is organized as follows: in the next section, Section 2, the electronic structure of beryllium chains is considered, showing how a linear disposition of the atoms can lead to relatively stable electronic states; in Section 3, the theoretical framework of the investigation is described; in Section 4, the optimized geometries, and the harmonic frequencies computed at this geometry, are presented and discussed; the singlet−triplet energy splitting is also investigated, showing that these chains have an antiferromagnetic behavior, with an exponentially decaying coupling constant as a function of the number of atoms; finally, in Section 5, some conclusions are drawn, and the problem of the existence and stability of these structures is addressed.

Table 1. Orbital Symmetry, Energies (hartree), and Occupation Numbers of the Occupied-Valence and Active CAS-SCF Orbitals for Be_7, at the Equilibrium Geometry[a]

orbital	symmetry	energy	occupation
1	$5\sigma_g$	−0.4123	2.000000
2	$4\sigma_u$	−0.4011	2.000000
3	$6\sigma_g$	−0.3765	2.000000
4	$5\sigma_u$	−0.3431	2.000000
5	$7\sigma_g$	−0.3024	2.000000
6	$6\sigma_u$	−0.2592	2.000000
7	$8\sigma_g$	−0.0738	1.004429
8	$7\sigma_u$	−0.0729	0.995571

[a] The orbitals are shown in Figure 1.

2. The Electronic Structure of Beryllium Chains

Beryllium is a metallic divalent element. The isolated atom has a closed-shell $1s^2 2s^2$ electronic configuration, so the

Table 2. Equilibrium Bond Distances (bohr) as a Function of the Chain Length N[a]

N	1	2	3	4	5	6
2	4.319					
3	4.115					
4	4.086	4.105				
5	4.094	4.107				
6	4.102	4.097	4.109			
7	4.106	4.099	4.110			
8	4.109	4.108	4.100	4.110		
9	4.111	4.108	4.100	4.111		
10	4.113	4.112	4.109	4.101	4.111	
11	4.114	4.113	4.109	4.101	4.111	
12	4.115	4.114	4.113	4.109	4.101	4.111
13	4.115	4.115	4.113	4.109	4.101	4.111

[a] Bond 1 indicates the central bond in the case of even N or the two central bonds for N odd and so on.

electronic structure of this element is somewhat reminiscent of a rare gas. For this reason, this species is not particularly reactive. In fact, beryllium has a relatively large Ionization Potential (IP), while the isolated atom does not admit any stable anionic state. In particular, beryllium IP is larger than the IP of the element immediately following in the same period (boron), a fact that is not a common feature on the Periodic Table. However, the empty $2p$ orbitals are close in energy to the occupied $2s$ one.[13] For this reason, and having two valence electrons only, beryllium is able to form at most two molecular bonds that are colinear. This is the case of the well-known species beryllium hydride (BeH_2) and the similar, although less studied, Be_nH_2 linear structures.[14]

Let us consider the case of a linear arrangement of N beryllium atoms, and let us take, for the sake of definiteness, the z axis as the chain direction. The symmetry group of the system is assumed to be $D_{\infty h}$ (of course, one has to verify this assumption in the course of the investigation). In particular, the orbitals will be labeled by g and u, according to their symmetry property with respect to the molecule-center of inversion.

A total of $N - 1$ Be−Be bonds are formed between neighboring atomic pairs. The bonds are the result of additive combinations of two sp_z hybrids, each one coming from one of the two atoms participating in the bond. These $N - 1$ bonds are able to host $2N - 2$ out of the $2N$ valence electrons of the system. The two remaining valence electrons can be placed in the last two valence σ orbitals, a σ_g and a σ_u. (The π valence orbitals are much higher in energy, and they do not play any important role in the qualitative description of the low-lying states at least for longer chains, as discussed in Computational Details section). They are obtained as linear combinations of the two outer sp_z hybrid bonds, one on each terminal beryllium atom. Because of their position in the chain, these external hybrids are called "edge", or "end", orbitals. If the chain is composed of several atoms, and considering that the optimum Be−Be distance in these structures is between four and five bohr, the two edge orbitals are quite separated, and therefore quasi degenerate. The lowest energy arrangement of the two electrons into the two orbitals is therefore, in a localized picture, one electron in each edge orbital. In terms of the delocalized symmetry orbitals, they correspond to the two configurations $(1)/(\sqrt{2})(\sigma_g\bar{\sigma}_g - \sigma_u\bar{\sigma}_u)$ (a singlet Σ_g) and $(1)/(\sqrt{2})(\sigma_g\bar{\sigma}_u + \sigma_u\bar{\sigma}_g)$ (a triplet Σ_u).

As far as the two edge orbitals, and the two corresponding electrons, are concerned, the situation is formally very much similar to the case of a stretched hydrogen molecule. Also in this case a singlet and a triplet form a manifold of degenerate states at long distance. The analogy can be pushed further, and in the case of beryllium chains one could consider the two low-lying ionic singlets arising from the double occupation of the edge orbitals, $(1)/(\sqrt{2})(\sigma_g\bar{\sigma}_u - \sigma_u\bar{\sigma}_g)$ and $(1)/(\sqrt{2})(\sigma_g\bar{\sigma}_g + \sigma_u\bar{\sigma}_u)$. However, these "ionic" states are likely to play an important role only in presence of a longitudinal electric field, and they will be the subject of future investigations.

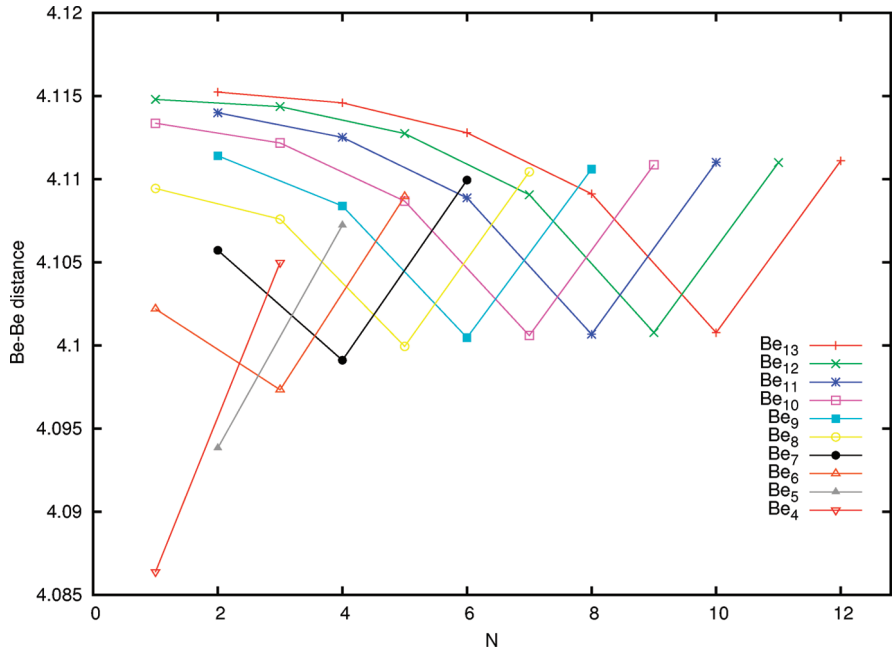


Figure 2. Bond lengths of different chains Be_N, from Be_4 to Be_{13}. Bond lengths in bohr.

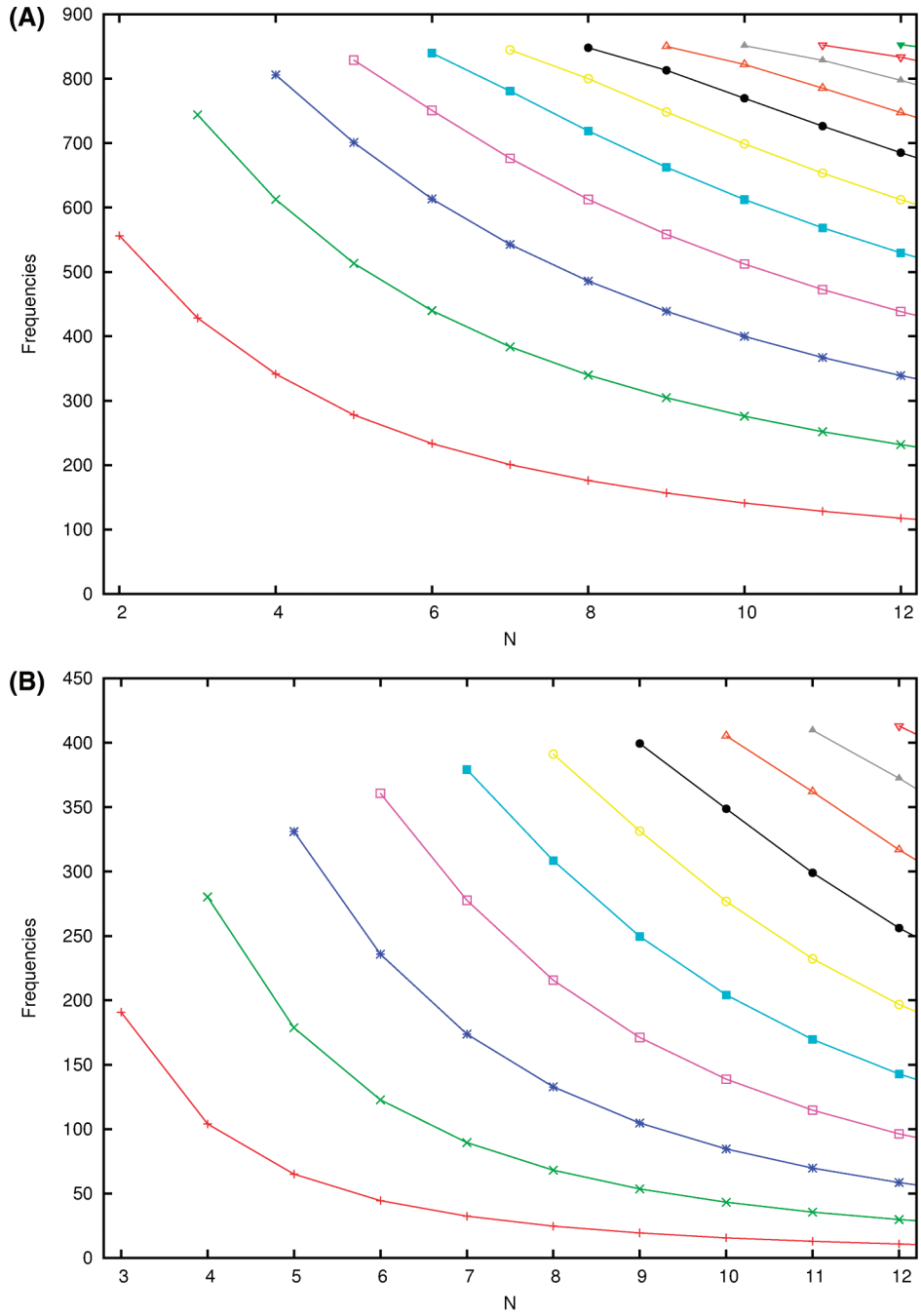


Figure 3. Harmonic frequencies for Be_N (each curve representing a different normal mode): stretching, from Be_2 to Be_{13} (a) and bending, from Be_3 to Be_{13} (b). Frequencies are in cm^{-1}.

3. Computational Details

In the course of the present investigation, a $D_{\infty h}$ symmetry has been assumed for the chain. However, for practical reasons, only abelian groups can be treated by most *ab initio* packages, so the actual calculation was performed in the D_{2h} abelian subgroup. Notice that, in principle, the linear chain could have a lower $C_{\infty v}$ symmetry. However, the fact that all the harmonic frequencies associated with the equilibrium geometry are real and positive excludes this possibility.

An ANO-type basis set was chosen as atomic orbitals, with the largest recommended contraction, $6s4p3d2f$.[15] Due to the multireference nature of the wave function at equilibrium (two Slater determinants having virtually identical weights), the SCF wave function is not able to give a correct description of the system. It is the CAS-SCF(2/2) description that gives the simplest qualitatively correct wave function. With this simple description we do not consider the influence of the π orbitals. The validity of this approximation can be assessed by analyzing the valence CAS-SCF of short chains. In particular the (per electron) π occupation goes from 0.06495 for the Be atom to 0.05469 for the dimer and 0.03199 for the tetramer. This fact clearly indicates a tendency toward a reduction of the π influence already evident in the dimer with respect to the isolated atom. Therefore, although π orbitals are crucial to correctly describe the dissociation of the Be_2, their influence becomes less important for longer chains, that are the main object of this investigation. The equilibrium geometries for the singlets and

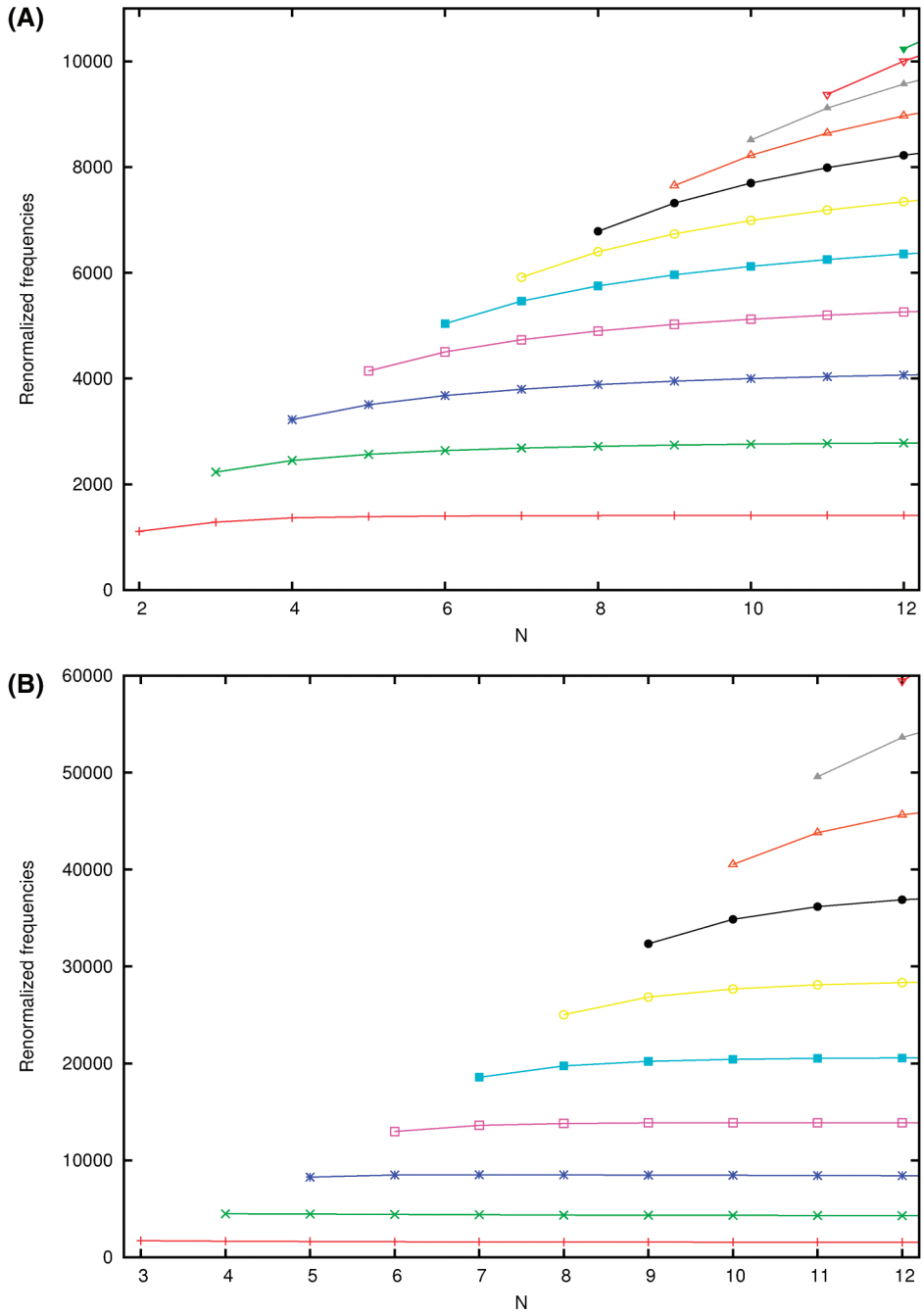


Figure 4. Renormalized harmonic frequencies for Be_N (each curve representing a different normal mode): stretching divided by N (a) and bendings divided by N^2 (b). Frequencies are in cm^{-1}.

the triplets, by imposing a linear constrain, were obtained at this level. The harmonic frequencies were then numerically computed at the equilibrium geometry for each chain. All the calculations presented in this study were obtained by using the MOLPRO[16] and DALTON[17] packages.

At the equilibrium geometry, singlet and triplet energies were obtained through a state-averaged CAS-SCF(2/2) calculation.[18−20] The state-averaged orbitals were then used in the CI calculations. The CI description introduces the effects of the dynamical correlation. Although, in particular for the longest chains, total energies computed by using this method suffer for the lack of size consistency, singlet−triplet splittings are certainly much less affected. The CI calculations reported in this work are of the internally contracted type[21,22] and were obtained by using the MOLPRO chain.[16]

In all CI calculations, the 1*s* orbitals were kept doubly occupied and frozen at the CAS-SCF(2/2) level.

4. Results and Discussion

We studied linear chains containing N beryllium atoms for several values of N. The behavior of these systems while permitting independent bond-length relaxations has been investigated. It turns out that, at equilibrium, the different bond lengths coincide within a few percent. In other words, the optimized geometries are very close to the equal-bond minimum geometries investigated in previous studies.

Table 3. CAS-SCF and MR-CI Absolute Energies (hartree) for the Singlet and Triplet State for Different Beryllium Chains, at the CAS-SCF Optimized Geometry[a]

N	MR-CI_{av} (T)	MR-CI_{av} (S)	CAS-SCF_{av} (T)	CAS-SCF_{av} (S)	CAS-SCF_{ss} (S)	RHF-SCF
2	−29.200974	−29.233715	−29.143642	−29.164091	−29.165722	−29.128752
3	−43.861856	−43.867080	−43.762741	−43.765410	−43.765580	−43.689933
4	−58.512805	−58.513587	−58.375529	−58.375899	−58.375916	−58.269425
5	−73.160347	−73.160466	−72.987399	−72.987451	−72.987453	−72.862037
6	−87.805814	−87.805836	−87.599010	−87.599018	−87.599019	−87.463211
7	−102.449702	−102.449709	−102.210526	−102.210527	−102.210527	−102.068880
8	−117.092274	−117.092278	−116.821995	−116.821995	−116.821995	−116.676827
9	−131.733710	−131.733713	−131.433440	−131.433440	−131.433440	−131.285896
10	−146.374143	−146.374145	−146.044870	−146.044870	−146.044870	−145.895630
11	−161.013680	−161.013683	−160.656291	−160.656291	−160.656291	−160.505732
12			−175.267707	−175.267707	−175.267707	−175.116102
13			−189.879119	−189.879119	−189.879119	−189.726639

[a] The geometry was optimized at a state-specific CAS-SCF level for the singlet (CAS-SCF_{ss}), while the MR-CI results were obtained after singlet-triplet averaged calculations (CAS-SCF_{av} and MR-CI_{av}).

In Figure 1, some of the CAS-SCF valence orbitals are shown, for a chain of intermediate length (Be_7) at the equilibrium geometry. (The seven doubly occupied core orbitals (four σ_g and three σ_u) have been omitted). The six doubly occupied canonical orbitals (Figure 1a) are different bonding combinations of Be−Be *sp* hybrid orbitals. The two active natural orbitals (Figure 1b) are the singly occupied edge orbitals. They are hybrid sp_z orbitals pointing toward the outer part of the chain. In Table 1, for this same system, the orbital energies and the occupation numbers of the valence doubly occupied or active orbitals are reported. The singly occupied orbitals have negative energies, although much higher than the energies of the occupied orbitals. Their occupation numbers are very close to 1.0, the $8\sigma_g$ orbital being slightly more occupied than the $7\sigma_u$ one. This is not surprising, if one considers the fact that they correspond to the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO), respectively. This confirms once more that these systems, at equilibrium, are biradicals whose electronic structure is very much reminiscent of the structure of an almost dissociated hydrogen molecule.

4.1. Optimized Geometries. In Table 2, the optimized bond lengths are reported. It appears that they depend in a very weak way on the bond position in the chain or the number of atoms in the chain. The general trends in bond lengths are more evident if one looks at Figure 2, where each curve is associated with a chain having a given value of N ($N = 2$ and $N = 3$ have been omitted). The external bond (labeled $N - 2$ in the figure) has a value of about 4.100 bohr. For all the chains, this is the longest one. This fact can be explained by the fact that the two terminal beryllium atoms are surrounded by only three electrons (instead of four, like all the remaining atoms placed in the inner region of the chain). The following bond ($N - 4$) is the shortest one for chains up to $N = 8$, and it is about 4.100 bohr. Then the bond lengths increase and saturate to a value close to 4.115 bohr. As a final remark, one can notice that the corresponding bond lengths in chains having different values of N are slightly shorter for the shortest chains.

4.2. Harmonic Frequencies. We computed numerical harmonic frequencies at the CAS-SCF(2/2) level, for both singlet and triplet states. It turns out that singlet and triplet have extremely close frequencies in the case of long chains, since these two states are almost degenerated. In other words, this means that the singlet−triplet quasi degeneracy is not removed by small bendings (or stretchings) of the system. For this reason, we report in this work only the frequencies for the singlet states. Harmonic vibrational frequencies for the triplet and the singlet states are reported in the Supporting Information, while a graphical representation of the frequencies is illustrated in Figure 3a (stretching modes) and 3b (bendings). The lowest-mode frequencies, both for bending and stretching, go to zero for large values of N, while the highest frequencies tend to finite values. As a general rule, stretchings are associated with higher frequencies than bendings, as the stretching of bond lengths is more energetic than the bending of bond angles. In particular, the high-frequency limit for the stretchings is almost exactly twice as much as the corresponding value for the bendings. It is interesting to notice that the lowest modes for stretchings and bendings follow the trends expected for a classical flexible rod in the harmonic approximation. In fact, the lowest longitudinal and transversal modes for such a system are approximatively given by

$$\omega_{long}(k, L) = \alpha_{long} k/L \tag{1}$$

and

$$\omega_{tran}(k, L) = \alpha_{tran}(2k + 1)^2/L^2 \tag{2}$$

respectively (see, for instance, refs 23 and 24). In eqs 1 and 2 L is the length of the rod, k is a positive integer, while α_{long} and α_{tran} are constants depending on the material and the shape of the rod. Therefore, the lowest stretching modes are inversely proportional to the rod length, while the lowest bending modes are inversely proportional to the square of the length. In Figure 4a,b, we have reported the frequencies of the stretching and bending modes divided by N and N^2, respectively. It appears that the lowest curves are essentially constants and follow in a remarkable way eqs 1 and 2. As expected, the accord worsens as the frequency increases, and it is completely lost in the case of the highest modes.

4.3. Energies and Singlet−Triplet Splitting. In Table 3, the total energies, obtained through the different methods, are reported. The nature of the interaction between the two localized edge orbitals depends on the value of the Heisenberg coupling J. This is defined as the difference between

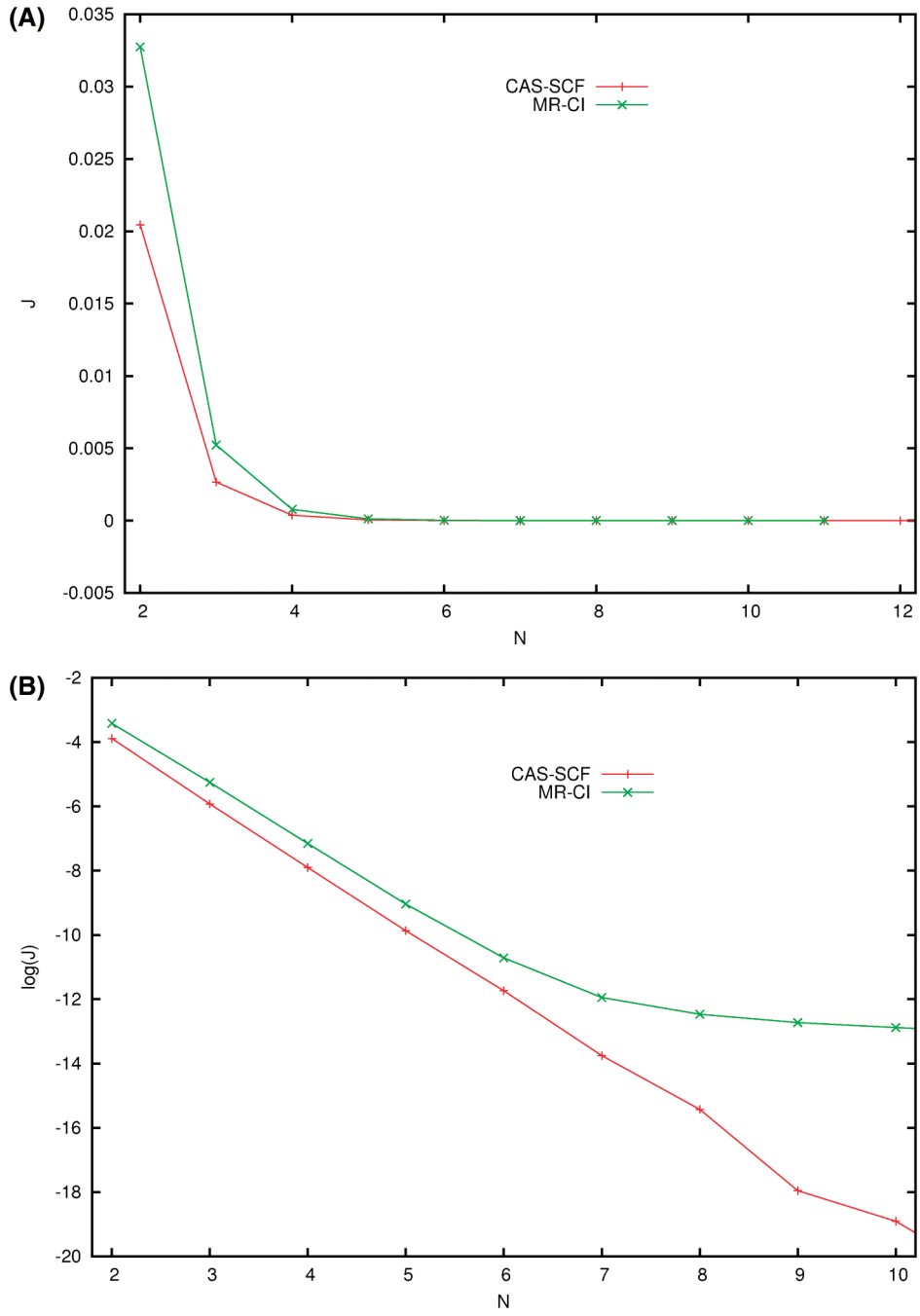


Figure 5. Singlet−triplet energy splitting (in hartree), $\Delta_{ST} = E_T - E_S$, as a function of N (a). Logarithm of the splitting (b). Red curves, CAS-SCF values; green curves, MR-CI values.

the triplet and the singlet energies. With this convention, a positive value of J corresponds to an antiferromagnetic coupling, as it happens for the chains we are discussing. A negative value of J, on the other hand, corresponds to a ferromagnetic coupling. The CAS-SCF and MR-CI spittings are shown in Figure 5a. It appears that the introduction of the dynamical correlation increases the value of the splitting by more than one-half with respect to the corresponding CAS-SCF values. In the region of small values of N, the CAS-SCF splitting goes exponentially to zero, as the number of beryllium atoms in the chain is increased, as shown in Figure 5b. In a logarithmic scale, the figure shows a linear behavior of J, both for CAS-SCF and MR-CI. The behavior of both MR-CI and CAS-SCF is linear up to $N = 6$, and the two curves are parallel up to this value of N. For long chains ($N > 9$), the CAS-SCF split become instable, being smaller that 10^{-8} hartree. The steep decrease of the (anti)ferromagnetic coupling with the distance of the center was already observed in other quasi-linear systems see for instance ref 25. It can be noticed from the logarithmic plot that the MR-CI split goes toward a constant value for chains longer than about ten atoms. This effect is most probably an artifact, and it could be due to the lack of size-consistency of the CI method. In fact, the energy split is severely underestimated at the CAS-SCF and MR-CI levels with respect to the FCI values,[12] while splits computed at a size-consistent level (NEVPT, see ref 26) do not present this behavior.[27]

5. Conclusions

We presented a theoretical CAS-SCF study on the equilibrium geometry and harmonic frequencies of the so far hypothetic Be_N linear chains. According to our results, beryllium linear chains composed of N atoms are local minima on the corresponding PESs. Although the effect of dynamical correlation was not included in the study, we think unlikely that this could qualitatively change this conclusion. For this reason, linear Be_N systems could be metastable clusters.

The equilibrium bond distances of the clusters depend in a very weak way on the number of atoms or the position of the bond in the chain. The harmonic frequencies follow the trends expected for the corresponding vibration of a classical bar. Our calculation were performed by using a harmonic approximation, so no values were obtained for the higher coefficients. However, the fact that analytical and numerical approaches gave virtually identical results is an indicator that anharmonic effects could be relatively small.

The singlet−triplet energy splitting can be described by an antiferromagnetic Heisenberg phenomenological Hamiltonian. The magnetic coupling follows an exponentially decreasing pattern as a function of N, which is roughly proportional to the distance between the unpaired edge electrons. The importance of the introduction of dynamical correlation for a correct description of the splitting was confirmed: For a fixed value of N, the CAS-SCF value of J is about two-thirds of the corresponding MR-CI result.

Although the linear forms are predicted to be local minima on the PESs, more compact clusters have quite lower energies. The possible existence of these clusters in linear geometries depends on the barrier height toward more stable arrangements. Linear chains could be formed by deposing Be atoms on an inert surface. Another possibility could be the stabilization through an interaction of a beryllium chain with some other species. In this contest, the possibility of hosting these chains within nanotubes is particularly attractive. Therefore, in our opinion, it is worth investigating under what circumstances (if any) these linear structures could be stable.

Acknowledgment. This work was partly supported by the French "Centre National de la Recherche Scientifique" (CNRS) and by the Italian Ministry of University and Research (MUR) and the University of Bologna under the project "PRIN 2006. Molecular Quantum Mechanics: Computational Methods and Analysis of Novel Phenomena". Support from the European Community under the COST D37 action is also gratefully acknowledged.

Supporting Information Available: Numerical values of computed harmonic frequencies for the various chains. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Emberly, E. G.; Kirczenow, G. *Phys. Rev. B* **1999**, *60*, 6028–6033.

(2) Bahn, S. R.; Jacobsen, K. W. *Phys. Rev. Lett.* **2001**, *87*, 266101.

(3) Himpsel, F. J.; Altmann, K. N.; Bennewitz, R.; Crain, J. N.; Kirakosian, A.; Lin, J.-L.; McChesney, J. L. *J. Phys.: Condens. Matter* **2001**, *13*, 11097–11113.

(4) Amorim, E. P.; da Silva, A. J. R.; Fazzio, A.; da Silva, E. Z. *Nanotechnology* **2007**, *18*, 145701/1–4.

(5) Crain, J. N.; Pierce, D. T. *Science* **2005**, *307*, 703.

(6) Davidson, S. G.; Stślicka, M. In *Basic Theory of Surface States*; Clarendon: Oxford, 1992.

(7) Hod, O.; Peralta, J. E.; Scuseria, G. E. *Phys. Rev. B* **2007**, *76*, 233401.

(8) Hod, O.; Barone, V.; Scuseria, G. E. *Phys. Rev. B* **2008**, *77*, 035411.

(9) Soon, Y.-W.; Cohen, M. L.; Louie, S. G. *Nature (London)* **2006**, *444*, 347.

(10) Monari, A.; Bendazzoli, G. L.; Evangelisti, S. *J. Chem. Phys.* **2008**, *129*, 134104.

(11) Monari, A.; Vetere, V.; Bendazzoli, G. L.; Evangelisti, S.; Paulus, B. *Chem. Phys. Lett.* **2008**, *465*, 102–105.

(12) Vetere, V.; Monari, A.; Scemama, A.; Bendazzoli, G. L.; Evangelisti, S. *J. Chem. Phys.* **2009**, *130*, 024301/1–9.

(13) Salomonson, S.; Warston, H.; Lindgren, I. *Phys. Rev. Lett.* **1996**, *76*, 3092.

(14) Tague, T. J.; Andrews, L. *J. Am. Chem. Soc.* **1993**, *115*, 12111–12116.

(15) Widmark, P.-O.; Malmqvist, P.-Å; Roos, B. O. *Theor. Chim. Acta* **1990**, *77*, 291.

(16) MOLPRO is a package of ab initio programs written by the following: Werner, H.-J., Knowles, P. J., with contribution from Almlöf, J., Amos, R. D., Berning, A., Deegan, M. J. O., Eckert, F., Elbert, S. T., Hampel, C., Lindh, R., Meyer, W., Nicklass, A., Peterson, K., Pitzer, R., Stone, A. J., Taylor, P. R., Mura, M. E., Pulay, P., Schütz, M., Stoll, H., Thorsteinsson, T., Cooper, D. L. (accessed March 10, 2009).

(17) DALTON, a molecular electronic structure program, Release 2.0 (2005) see:http://www.kjemi.uio.no/software/dalton/dalton.html (accessed March 10, 2009).

(18) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157.

(19) Knowles, P. J.; Werner, H.-J. *J. Chem. Phys.* **1985**, *82*, 5053.

(20) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259.

(21) Knowles, P. J.; Werner, H.-J. *J. Chem. Phys.* **1988**, *89*, 5803.

(22) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1988**, *145*, 514.

(23) Landau, L.; Lifschitz, E. In *Course Of Theoretical Physics*, 2nd ed.; Pergamon: Vol *7*: Theory Of Elasticity, p 116.

(24) Moloney, M. J. A. *J. Phys.* **2005**, *73*, 439–441.

(25) Peng, S.-M.; Wang, C.-C.; Jang, Y.-L.; Chen, Y.-H.; Lin, F.-Y.; Mou, C.-Y.; Leung, M.-K. *J. Magn. Magn. Mater.* **2000**, *209*, 80–83.

(26) Angeli, C.; Pastore, M.; Cimiraglia, R. *Theor. Chem. Acc.* **2007**, *117*, 743–754.

(27) Pastore, M.; Monari, A.; Bendazzoli, G. L.; Cimiraglia, R.; Evangelisti, S. submitted to *J. Chem. Phys.*

CT9001027

JCTC Journal of Chemical Theory and Computation

Confined But-2-ene Catalytic Isomerization Inside H-ZSM-5 Models: A DFT Study

Giampaolo Barone,[†] Nerina Armata,[†] Antonio Prestianni,[†] Teresa Rubino,[†] Dario Duca,*[,†] and Dmitry Yu. Murzin[‡]

Dipartimento di Chimica Inorganica e Analitica "S. Cannizzaro" dell'Università, viale delle Scienze Ed. 17, I-90128 Palermo, Italy, and Laboratory of Industrial Chemistry, Process Chemistry Centre, Åbo Akademi University, Biskopsgatan 8, FIN-20500 Åbo/Turku, Finland

Received September 25, 2008

Abstract: The isomerization of *cis*-but-2-ene to *trans*-but-2-ene within a 22T H-ZSM-5 zeolite model, also in the presence of two adsorbed Pd atoms, has been studied by DFT calculations. The results obtained allow us to state that the *cis*/*trans* but-2-ene isomerization can easily proceed inside unsupported zeolite cavities. In this case, differently than in the gas phase reaction, the *trans*-but-2-ene is less stable than the *cis*-but-2-ene, when adsorbed on the zeolite inner surface. Excluding the adsorption−desorption steps, the isomerization process involves two intermediates and three transition states, whose energy content is always very low with respect to that of reagents and intermediate species. The reaction is in principle allowed also in the presence of two Pd atoms embedded inside the zeolite cavity. However, strong H-Pd interactions seem to cause higher activation energies along the formation of the involved intermediates and transition states. To evaluate the confining effects of the zeolite room on the *cis*/*trans* isomerization process, the latter has been also analyzed on protonated (Pd_2H^+) and unprotonated (Pd_2) bare palladium fragments at different multiplicity states. The but-2-ene adsorption on the considered systems and the mutual influence occurring between the metal atoms and the hydrogen acidic sites at different multiplicity states have also been taken into consideration.

1. Introduction

Modeling of catalyzed processes is extensively employed to check and design reaction environments and mechanisms involved in heterogeneous catalytic systems.[1] Also the zeolite properties have been thoroughly investigated by computational approaches.[2,3]

Zeolite-based materials are extensively used in the catalytic hydrocarbon transformation, namely in cracking, hydrocracking, and C5−C7 isomerization.[4] The use of the acidic solid catalytic media, such as zeolite materials, with respect to the liquid ones, has indeed important technological and environmental implications related to selectivity, substrate separation, catalyst regeneration, large operating temperature ranges, absence of toxic byproduct formation, and, in general, facility of use, also avoiding waste production.[5] In particular, the ZSM-5 derived catalysts are used in fuel-cracking and isomerization as well as in methanol-to-gasoline conversion.[6]

Of course, the morphological and the chemical characteristics of the zeolite materials play an important role in the molecular surface adsorption and in the catalytic activity and selectivity.[7,8] In particular, the acidic properties caused by the substitution of Si^{4+} ions by couples of Al^{3+} and H^+ ions are fundamental in determining the local electrophilic character of the zeolite framework that is involved in driving the cation-promoted catalytic reactions in confined space.[9−11] In fact, the acidic Al-substituted zeolite locally provides a protonic environment to the adsorbed catalytic substrate (*Ad*), originating carbocation (carbonium and/or carbenium) intermediate (*In*), and transition state (TS) species,[12,13] which, mainly for the small olefin derivatives trapped inside the

* Corresponding author e-mail: dduca@cccp.unipa.it.
[†] Università di Palermo.
[‡] Åbo Akademi.

10.1021/ct800402k CCC: $40.75
Published on Web 04/21/2009

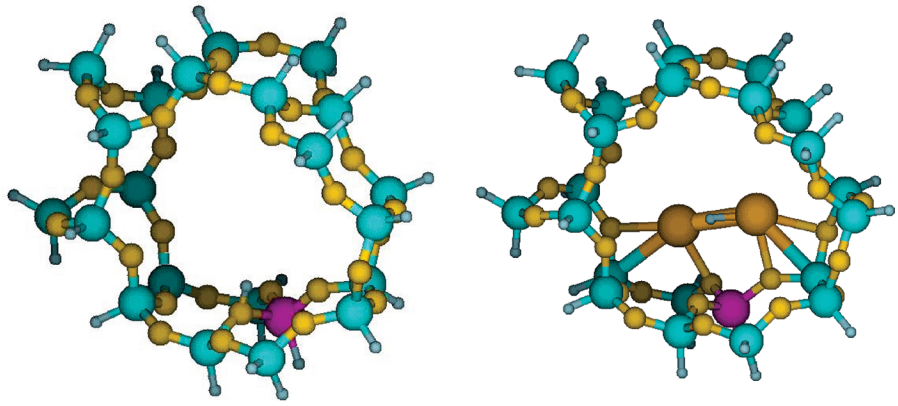


Figure 1. Singlet state HLBs optimized 22T H-ZSM-5 and Pd_2/H-ZSM-5 fragments employed in the kinetic study: H atoms, gray-blue; O atoms, gold; Si atoms, cyan; Al atom, purple; Pd atoms, brownish. The E_0 values for (a) and (b) are equal to −8598.7205 and −8852.2354 au.

zeolite cavities, are bulk-alkoxide and framework-anchored carbenium species.[14]

It is widely accepted that the computational chemistry has an enormous impact on the development of the zeolite catalytic materials and processes.[3] Focusing on the ZSM-5 systems, a few bibliographic examples of interest for the present work are reported in the following. Specifically, they concern i) reactivity,[13,15] ii) adsorption−desorption processes,[8,11] iii) proton transfer in the presence of metal centers,[16,17] and iv) Si/Al molar ratio effects and characterization of acidic sites.[18,19]

However, the employed models' sizes are not always adequate.[13,15,19,20] In this case, the lack of the zeolite cavity framework in studying the catalytic processes, hence the absence of local interactions between the catalytic substrate and the inner wall of the catalyst, generally causes an overestimation of the involved activation energy[21] and an unreliable estimate of the relative stability of molecules participating in a given reaction,[13] in particular, flattening the energetics of the involved zeolite systems.[19]

Macromolecular 22T H-ZSM-5 zeolite cluster models, also containing palladium centers, have been employed in the present work to study the *cis*/*trans* but-2-ene isomerization inside the zeolite cavity. Hydrogenation,[22] dehydrogenation, and double-bond migration[23] as well as oligomerization of the butene molecules[24] or skeletal transformations that involve the 2-methylpropene formation[25] were not considered, being beyond the scope of the present work. In fact, the *cis*/*trans* title isomerization is rooted in pioneering works on photo- and iodine-catalyzed *cis*- to *trans*-but-2-ene isomerization,[26] and it is up till now considered as an introductory example in understanding the basic principles of catalysis.[27,28] The choice of studying the *cis*/*trans* isomerization process inside the zeolite cavity in the presence of small metal clusters is conversely desirable because isomerization catalysts (e.g., metals) and shape selective materials (e.g., zeolites), taken together, can be effective in butene double bond migration/isomerization processes of practical interest.[29]

This paper, referring to uncatalyzed as well as to bare-palladium catalyzed processes, examines the confinement effects of the H-ZSM-5 zeolite cavity and the influence of embedded palladium atoms on the proton-activated *cis*/*trans* but-2-ene isomerization. Electronic charge transfers involved in the isomerization processes inside the zeolite cavity embedding or not the palladium atoms are also qualitatively investigated.

The paper organization is explained as follows. In section 2, the properties of the models employed are given with the essential features characterizing the computational methods. In section 3, details on the structural and mechanistic findings are presented, discussing the following: the metal effects on the zeolite Brønsted acidity sites (*BASs*) (section 3.1); the but-2-ene adsorption on zeolite fragments (section 3.2); the uncatalyzed but-2-ene isomerization (section 3.3); and the catalyzed but-2-ene isomerization on the zeolite fragments (section 3.4). Supporting Information is available and concerns the following: the influence of different basis sets employed to study the title reaction; the aggregation sites of the palladium atoms inside the zeolite cavity; and the multiplicity states of the supported systems.

2. Computational Details

Structural and kinetic properties of but-2-ene derivatives, both isolated and within a 22T H-ZSM-5 zeolite model, were calculated. The 22T H-ZSM-5 system, also supporting palladium atoms (Pd_2/H-ZSM-5), consists of two 10T rings connected by two T moieties (see Figure 1). Reference calculations on the same hydrocarbons, interacting with bare palladium fragments, both protonated (Pd_2H^+) or not (Pd_2), were also performed. Different multiplicity states were taken into consideration either for the bare-palladium or for the supported zeolite systems. The zeolite cavity chosen is a model of the largest pore present in a ZSM-5 zeolite. As recently reported for a very similar reaction system,[15] the cluster approach has been employed.[30] The starting zeolite model (ZSM-5), whose coordinates were taken from a DLS refined X-ray structure,[31] consists of 81 atoms of which 28 cutoff (capping) hydrogens, bound to silicon atoms, substitute the original oxygens to truncate the periodic zeolite structure. In order to mimic the acidic zeolite system (H-ZSM-5), one

silicon was substituted by one aluminum atom whereas one hydrogen atom was bound to a T_4OT_1 position,[32] through an oxygen site vicinal to the aluminum and pointing within the zeolite cavity[19] (see Figure 1a). This hydrogen acidic site corresponds to one of the more stable *BASs* in the largest zeolite H-ZSM-5 10T ring,[19] and for this reason it has been selected to mimic a possible catalytic site within a zeolite cavity.

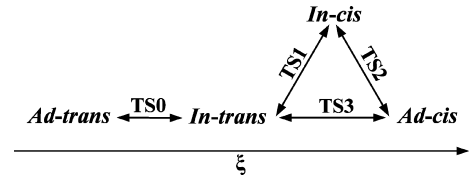
The calculations were performed using the Gaussian03[33] program package, by optimizing the different geometries at the DFT B3LYP[34] level. This choice was supported by the results reported by Tirado-Rives and Jorgensen,[35] which showed that this DFT functional is very effective in estimating the isomerization of several organic molecules. Time-dependent[36] and unrestricted[37] DFT calculations were performed on the species involved in the *cis*/*trans* uncatalyzed isomerization.

Two different groups of basis sets, in the following indicated as low level basis sets (LLBs) and high level basis sets (HLBs), were mainly employed for the calculations. The LLBs procedure makes use of i) the 6-31G(d,p)[38] basis set for the atoms of the but-2-ene derivatives involved in the catalyzed reactions (section 3.4) and for the Al, O, and the acidic H atoms of the aluminum-modified T environment, ii) the 3-21G[39] basis set for the Si, O, and H atoms of the remaining 21T fragments, and iii) the LANL2DZ[40] basis set for the Pd atoms. The HLBs approach differs from the LLBs one due to the use of the 6-31G(d,p) basis sets for the silicon and oxygen atoms belonging to the 21T moieties of the zeolite rings.

The uncatalyzed (section 3.3) and the bare-palladium catalyzed (section 3.4.2) reactions were investigated by the HLBs. The geometries at the energy minima were optimized by employing both HLBs and LLBs. Moreover, single point calculations (SP_{df}) were performed on structures optimized by using the HLBs and employing the 6-311+G(2d,2p)[41] instead of the 6-31G(d,p) basis set. The LLBs was chosen to limit the computational time with the hypothesis that the differential energetics and the local properties of the treated zeolite systems could be described with adequate accuracy even with a small basis set. The HLBs and the SP_{df} approaches were conversely employed to test the reliability of the hypothesis above.

Vibration analysis within the harmonic approximation of the energy minimum structures was performed[42] on all the optimized geometries. Due to the fact that the geometry of the investigated models was partially optimized, fixing the coordinates of the cutoff external hydrogens, 29−31 spurious imaginary frequencies were always observed, essentially identical in all the systems. This suggests that an error cancelation should occur when considering relative values of the calculated standard free energy and enthalpy at a given temperature. Both the thermochemical parameters above were calculated at 298.15 K to study the acidic properties of the considered H-ZSM-5 systems, after the zero-point energy correction,[42] always in the frame of the mentioned harmonic approximation of the vibrational potential. Transition states, calculated by the LLBs approach, were fixed by the synchronous transit-guided quasi-Newton method.[43] Noticeably, during the vibration analysis of the TSs, it was always possible to find one imaginary frequency attributable to a corresponding reaction coordinate.[42] As a consequence, we have been able to discriminate among energy-minimum and saddle-point constrained geometries. Finally, in the evaluation of the adsorption energies, characterizing the interaction of the hydrocarbon and the 22T fragments, the BSSE correction[44] was always introduced, namely for the SP_{df}, HLBs, and LLBs systems.

Scheme 1. Details on the *cis*/*trans*-But-2-ene Isomerization Pathway



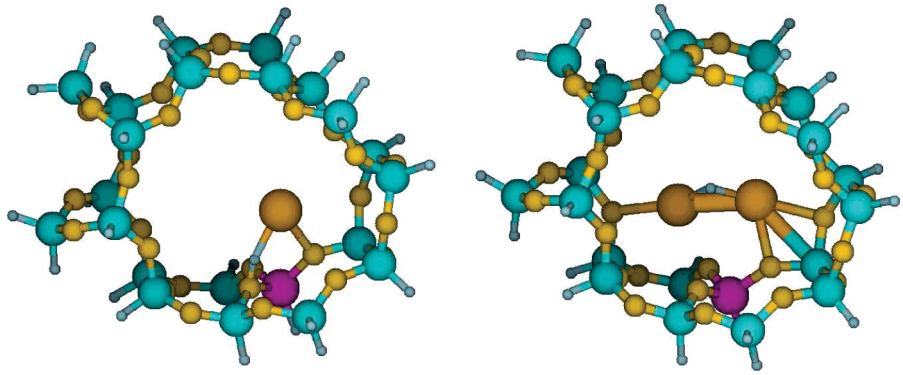
Starting from the X-ray coordinates of the ZSM-5 model, partial geometry optimizations were initially performed on the 28 external cutoff hydrogen atoms; subsequently, by fixing the coordinates of the external hydrogen atoms, the partial optimization was extended to the remaining atoms of the 22T unsupported (H-ZSM-5) and supported (Pd_2/H-ZSM-5) models. Figure 1 shows the optimized geometries of H-ZSM-5 and Pd_2/H-ZSM-5, being the corresponding structures obtained by HLBs and LLBs indistinguishable by sight.

3. Results and Discussion

The calculated *cis*/*trans* but-2-ene isomerization pathway, occurring inside the H-ZSM-5 zeolite cavities, is shown in Scheme 1. The reaction-coordinate (ξ) connects the least to the most stable but-2-ene isomer found inside the zeolite fragments, namely the *trans*-but-2-ene to the *cis*-but-2-ene (see section 3.4). *Ad-cis* and *Ad-trans* represent the corresponding but-2-ene adsorbed species, while *In-cis* and *In-trans* represent the intermediate conformer species geometrically more similar to the *Ad-cis* and *Ad-trans*, respectively. Finally, TS0-TS3 are the transition states along the catalytic isomerization pathway, with TS3 obtained only for the unsupported zeolite model. It is worth pointing out that both structures and energies of TS2 and TS3 were essentially coincident.

The reaction mechanism sketched in Scheme 1 is only a portion of the whole catalytic butene transformation process that also implicates by side surface reactions[22−26] as well as adsorption, desorption, and diffusion of the involved species. In the present paper we have focused on the effects of the zeolite confinement and on the influence of embedded palladium atoms on the reaction steps, leading from *cis*- to *trans*-but-2-ene and *vice versa*. The nonactivated adsorption energy (E_{ad}) values of the *cis* and *trans* species are discussed in section 3.2.

3.1. Metal Effects on the H-ZSM-5 Acidity. A preliminary study (see the Supporting Information) points out that the catalytic activity of the bifunctional palladium supported H-ZSM-5 systems toward the hydrogen-induced reactivity of light unsaturated hydrocarbons could be mainly attributed



(a) (b)

Figure 2. Palladium supported systems: (a) singlet ground-state HLBs optimized 22T Pd/H-ZSM-5 fragment: $E_0 = -8725.4620$ au and (b) triplet ground-state HLBs optimized 22T Pd_2/H-ZSM-5 fragment: $E_0 = -8852.2354$ au.

Table 1. Deprotonation of Different H-ZSM-5 Fragments: Values of Energy and Standard Free Energy at 298.15 K Calculated at the Ground States by the LLBs and HLBs Approaches

	LLBs		HLBs	
fragment[a]	ΔE/kJ/mol	$\Delta G°$/kJ/mol	ΔE/kJ/mol	$\Delta G°$/kJ/mol
H-ZSM-5	1278	1252	1254	1224
Pd/H-ZSM-5	1297	1272	1277	1250
Pd_2/H-ZSM-5	1448	1419	1378	1362

[a] Both protonated and nonprotonated H-ZSM-5 and Pd/H-ZSM-5 fragments showed singlet ground states, whereas the Pd_2/H-ZSM-5 fragment was at triplet and singlet state in the protonated and nonprotonated forms, respectively.

to the availability of (ionic or atomic) hydrogens -- either derived from *BASs* and/or produced by external sources -- interacting with small palladium fragments adsorbed nearby the aluminum-modified T fragments.

Several palladium supported fragments have been considered (see the Supporting Information) with the general formula Pd_n/H-ZSM-5 (with $n = 1-2$). The analysis of the most stable systems found points out that the acidic hydrogens near the aluminum atom represent the favorite nucleation centers for the deposition of palladium atoms, see Figures 1b and 2. This result induced us to evaluate the relative acidity of the most stable Pd_n/H-ZSM-5 model systems, presuming that local acidity changes caused by the palladium atoms embedded in the zeolite framework could affect the catalytic activity. The increased stability of the metal supported systems is indeed an interesting issue related to the decrease of the zeolite acidity strength, detected experimentally in the presence of adsorbed Group 10 metal crystallites.[45]

The acidity strength was estimated by calculating the values of deprotonation energy[10] and standard free energy at 298.15 K. Table 1 shows the results concerning the unsupported and supported fragments: H-ZSM-5, Pd/H-ZSM-5, and Pd_2/H-ZSM-5. The energy and standard free energies, irrespective of the basis set level, show that the strong acidity of the zeolite is decreased by the presence of also low metal content, in agreement with the experiment.[45] Noticeably, the deprotonation energy values of the nonsupported fragment calculated by the LLBs and HLBs procedures are in good agreement (ca. 5% and 3% of difference, respectively) with the analogous parameter calculated for H-ferrierite by Nieminen et al.[10]

The decrease of the acidity of the H-ZSM-5 fragments with the increase of the number of adsorbed Pd atoms can be ascribed to hydrogen migration from the support to the embedded metal atoms. Hence, the acidity, if referred to the mobility of H from the -OH acidic group, is locally increased by the metal atoms that, however, fix the same hydrogen, producing at a macroscopic level an apparently lower acidity. There is indeed a strong affinity between the H species and the Pd_2 fragment, such that the character of the former changes from acidic to hydride-like. This behavior is supported by a Mulliken population analysis (not shown), from which electron deficency on the metal atoms can be deduced, in agreement with experimental findings.[45]

On the whole, our results nicely explain the modified acidic properties of the bifunctional zeolite derivatives, whose experimental catalytic properties already seemed to be affected by the mutual interaction of the metallic and acidic functions.[45] This further justifies the choice to investigate the influence of the palladium atoms in the but-2-ene isomerization inside the H-ZSM-5 fragment. As already mentioned, the Pd_2/H-ZSM-5 system was employed for this purpose. Since this one as well as the smaller reference systems -- Pd_2 and Pd_2H^+ -- showed singlet ground state when interacting with hydrocarbon derivatives (see section 3.4.2), the kinetic study was followed just in the singlet state. Moreover, due to the close similitude found, by applying the HLBs and LLBs approaches, on the geometrical and the energetic parameters of both the supported and the nonsupported H-ZSM-5 catalyst models (see the Supporting Information), we decided to detail the title kinetic study by the less time-consuming method, namely the LLBs one.

3.2. But-2-ene Adsorption on the H-ZSM-5 Fragments. In Table 2 the BSSE corrected adsorption energy of *cis*-but-2-ene and *trans*-but-2-ene on the H-ZSM-5 22T fragments embedding or not palladium atoms is reported, using different basis sets. It is also reported the coefficient of variation calculated for the counterpoise BSSE corrections corresponding to the energy minimum species involved in the reaction paths of Scheme 1. The unsupported H-ZSM-5 systems showed that the but-2-ene adsorptions on the T_4OT_1 site[15] occur via π interactions. The LLBs and HLBs adsorp-

Table 2. BSSE Corrected Ground-State Adsorption Energy of *cis*-But-2-ene and *trans*-But-2-ene on H-ZSM-5 22T Fragments without and with Palladium, Using Different Basis Sets

approach level	Ead_{cis}[a]/kJ/mol	Ead_{trans}[a]/kJ/mol	vc[a,b]
LLBs	22.3 | 190.3	13.2 | 182.2	0.003 | 0.036
HLBs	14.5 | 162.5	4.7 | 141.2	0.046 | 0.012
SP_{df}	12.6 | 200.2	3.4 | 167.8	0.055 | 0.012

[a] On the left and on the right, the values of the unsupported and supported systems are reported, respectively. As the H-ZSM-5 system, both the adsorbed fragments were in singlet ground state, while the Pd_2/H-ZSM-5 ground state was a triplet. [b] Coefficient of variation -- $vc = (\sigma)/(\mu)$, with σ and μ being the standard deviation and the mean values calculated for the counterpoise BSSE correction values, respectively. These are determined by a given basis set, for all the energy minimum species involved in the reaction paths of Scheme 1.

tion energies are in the range of those calculated for larger 52T H-ZSM-5 fragments by DFT B3LYP cluster approaches.[19]

An inversion of the stability order between *cis* and *trans* adsorbed species is observed, both for the embedding and nonembedding palladium atoms fragment and irrespective of the basis set considered. This result, which is in agreement with the activation energies characterizing the diffusion of the same but-2-ene isomers in analogous unsupported zeolite systems,[46] could be explained considering the constraining effects of the zeolite cavity[10,19,46] and in particular the lower steric repulsion felt by the *cis*-but-2-ene isomer within the 10T zeolite cavity. For the unsupported systems a characteristic decreasing trend of the adsorption energies, with the increasing of the basis set size, is clear. The same trend is conversely not observed for the Pd supported fragments.

In fact, it has not been yet possible to determine the experimental adsorption energies of these systems,[46] and in a recent study on the 2-methylpropene on unsupported H-ferrierite, Tuma and Sauer turned to the consideration of the corresponding saturated hydrocarbon adsorption energies in order to have experimental references.[11] The adsorption energies obtained by those authors using a hybrid MP2:DFT approach accounting for the dispersion effects occurring inside the zeolite cavity showed even higher values than the experimental references. Analogous results were reported for similar systems, using QM/MM methods[10] designed to include dispersion effects.

The adsorption energy values of Table 2 are smaller than those obtained by approaches specifically designed to include dispersion effects[10,11] while, as a paradox, becoming even smaller increasing the calculation level, up to include diffuse functions in the DFT B3LYP treatments. On the other hand, the results reported in Table 2 are in good agreement with DFT periodic calculations performed at a comparable level.[11] However, it can be also considered that the difficulties to perform experimental measurements of the π adsorbed hydrocarbons on -OH groups of acidic zeolites might be a consequence of the high reactivity of the resulting adduct.[46]

Whereupon, in our opinion, it is not straightforward to conclude that zeolite cavities containing highly reactive unsaturated hydrocarbons must show either higher or lower local energies, compared to those zeolite cavities containing less reactive saturated hydrocarbons. In any case, the values of the *vc* parameter reported in Table 2 clearly show that the variations on the average BSSE correction value with respect to a given basis set are very small irrespective of the involved zeolite system (i.e., supported or nonsupported). As a consequence, in agreement with the recent works of Gleeson[15] and Tirado-Rives and Jorgensen,[35] it is reasonable to consider that the results of the energetic differential approach to the study of the title isomerization -- and in particular of the steric pressure exerted by the cavity on the reaction species -- has to be considered consistent also when using the LLBs.

Furthermore, it is to be observed that the singlet ground-state HLBs di-σ adsorption energy of the *cis*-but-2-ene resulted 202.0 and 115.0 kJ/mol on the bare Pd_2H^+ and Pd_2 fragments, whereas in the same order on the same fragments the di-σ adsorption energy was 199.0 and 117.6 kJ/mol for the *trans*-but-2-ene. The unprotonated and the protonated Pd_2 bare fragments showed singlet and triplet ground states, whereas the di-σ adsorption mode, in agreement with Delbecq and Zaera,[28] always resulted in being more stable than the π mode in the ground state, irrespective of the considered but-2-ene isomer. Finally the relative stabilization of the *cis* species with respect to the *trans* one in the Pd_2H^+ systems is stressed.

Comparing these findings with those summarized in the row 6-31G(d,p) of Table 2, it is straightforward to recognize the role, in the but-2-ene adsorption processes, of both the acidic center (stabilizing) and of the zeolite cavity (destabilizing). Of course, these properties will affect catalytic reactivity and selectivity of the zeolite material.

3.3. Uncatalyzed But-2-ene Isomerization. The *in vacuo* *cis*/*trans* but-2-ene isomerization should involve i) the C2-C3 lengthening (double-bond breaking), ii) the rotation of 180° around the C2-C3 bond, and iii) the C2-C3 double-bond back-formation. Similar transformations characterize the rotation around the double bond of ethylene. Post HF calculations revealed the existence of a possible photochemical mechanism in which the TS of ethylene, showing a 90° rotation around the C-C double bond, would be a degenerate singlet−triplet spin state, $^1(\pi\pi^*)$ and $^3(\pi\pi^*)$ respectively, characterized by a relative energy of ca. 300 kJ/mol higher than the ground state.[47]

By using unrestricted DFT calculations,[37] we found a $^3(\pi\pi^*)$ state with the C2-C3 rotated by 90° that can be identified as the TS species in the gas phase *cis*/*trans* isomerization. This species shows an energy value 260 kJ/mol higher than that of the most stable isomer (*trans*-but-2-ene). The value is in good agreement with the one reported for ethylene and also with the experimental value found in the gas phase *cis*/*trans* isomerization (264 kJ/mol).[27] Further calculations on the uncatalyzed systems showed that i) the *trans*- and *cis*-but-2-ene ground state differ by ca. 5 kJ/mol, being the first more stable, ii) the Franck−Condon $\pi^2 \rightarrow {}^1(\pi\pi^*)$ *trans*-but-2-ene transition energy determined at the TD-DFT level[36] is equal to 737 kJ/mol, in excellent agreement with the value found for the ethylene system (800 kJ/mol),[47] and iii) both the unrotated *cis*- and *trans*-but-2-ene $^1(\pi\pi^*)$ states are less stable than the unrotated *trans*-

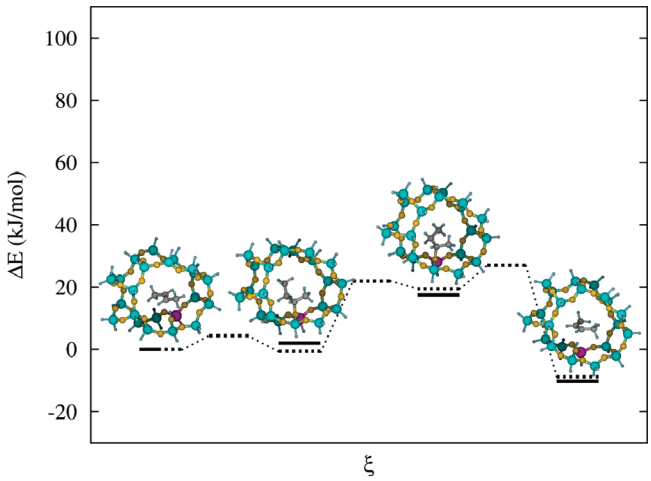


Figure 3. Catalytic pathway for the but-2-ene isomerization in the 22T H-ZSM-5 system and optimized ground-state HLBs minima: from the left, the reactants (*Ad-trans* and *Ad-cis*) and the intermediates (*In-trans* and *In-cis*) are shown. The *Ad-cis* species is the most stable fragment in the path: in the LLBs calculation, dotted bars, $E_0 = -8712.9048$ au, in the HLBs calculation, solid bars, $E_0 = -8755.9702$ au. Following the relative stability of the but-2-ene isomers inside the H-ZSM-5 fragment, the reaction coordinate, ξ, goes from the left to the right.

but-2-ene $^3(\pi\pi^*)$ state by ca. 300 kJ/mol. Hence, identifying the TS occurring in the *in vacuo cis*- to *trans*-but-2-ene isomerization with states like the relatively more stable $^3(\pi\pi^*)$, having the C2-C3 rotated by 90°, we can rationalize the found low interconversion speed by the high activation energies characterizing the involved TS.

3.4. Catalyzed But-2-ene Isomerization. *3.4.1. Unsupported H-ZSM-5 Fragment.* The *cis*/*trans* but-2-ene isomerization was simulated within the H-ZSM-5 cavity shown in Figure 1a. At variance with the π-bonded *cis*- and *trans*-but-2-ene, in the structures of the intermediates (*In-trans* and *In-cis*) and transition states (TS0, TS1, TS2, and TS3) the zeolite acidic hydrogen binds (moves toward) one of the carbon atoms -- by convention C3 -- involved in the alkene bond, reducing the double to a single bond, while the other central carbon atom (C2) is bound to one oxygen (usually that originating from the acidic -OH) of the aluminum-modified T moiety. For these reasons the isomer interconversion becomes easier, with respect to the uncatalyzed process, and occurs in the ground state by hydrogen addition−elimination.

Ignoring for the sake of simplicity the reaction step including TS3, a low energy kinetic path for the *cis*/*trans* title isomerization inside the zeolite cavity is shown in Figure 3: besides the *Ad-trans* and *Ad-cis* species, two *In*-states and three TSs, with LLBs energy barriers below 25 kJ/mol, are involved. Clearly, the energy minimum structures and the corresponding differential energetics obtained by the LLBs and the HLBs procedures are very similar to each other.

Interestingly, although the involved relative energies are of the same order of magnitude as those reported by Gleeson,[15] it is remarked here that we have found two reaction mechanisms in agreement with Scheme 1. The first, as the one reported by Gleeson, involves one intermediate and two transition states (TS0 and TS3 ≡ TS2), while the second involves two intermediates and three transition states (TS0, TS1, and TS2).

In the pathway of Figure 3 the nonactivated alkene adsorption processes are not explicitly reported; by including them the number of the involved elementary events would correspond to those reported for the iodine catalyzed process.[27] Noticeably, the iodine and the zeolite catalyzed rotation mechanisms and energetics are very similar, and their activation energy barriers are both significantly lowered, compared to the gas phase. The energy drop in both catalytic mechanisms is mostly related to the (sp^2 to sp^3) character change of the C2-C3 alkene bond. However, the *cis*/*trans* zeolite isomerization process could be more favored, with respect to that catalyzed by iodine, in the absence of catalyst aging caused by the parallel carbonaceous formation on the zeolite surface.[25] Indeed, instead of the activated iodine dissociation implying ca. 75 kJ/mol,[27] the catalytic process on the acidic zeolite starts by the nonactivated alkene adsorption and moreover, due to the confinement effects of the zeolite cavities, the *Ad-cis* becomes more stable (ca. 9 kJ/mol at the LLBs level) than the *Ad-trans* derivative.

3.4.2. Supported Pd_2/H-ZSM-5 Fragment. The consideration of the bifunctional metal supported systems was mainly addressed to analyze the effects of the reduced room inside the zeolite cavity and the influence of the modified local zeolite acidity for the presence of metal atoms.[16,17,45] For this reason also the bare-palladium fragments were considered. The singlet states always resulted in being the most stable in the Pd_2, Pd_2H^+, and Pd_2/H-ZSM-5 systems when adsorbing hydrocarbon species. The Pd_2 and Pd_2/H-ZSM-5 systems if not adsorbing hydrocarbons, in agreement with the data of Moc et al.,[48] conversely resulted in the triplet ground state, whereas the protonated Pd_2H^+ species showed a singlet ground state (for details on zeolites embedding Pd atoms, see the Supporting Information).

As already pointed out, all the species involved in the reaction path showed a singlet ground state: as an example the *Ad-trans* and *Ad-cis* triplet states were less stable than the corresponding singlet state species by ca. 40 and 150 kJ/mol, irrespective of the basis set employed. The quintuplet states were also less stable and the corresponding geometries typically lost, as previously observed for the but-2-ene derivatives adsorbed on the Pd_2 and Pd_2H^+ bare fragments, the di-σ adsorption modes shown in the more stable multiplicity states.

The supported zeolite fragment of Figure 1b, obtained by optimizing the Pd_2/H-ZSM-5 system, shows the result of a "reverse spillover" mechanism,[16,17,49,50] namely the diffusion of the acidic hydrogen and its adsorption in between the two Pd atoms. A possible reaction path of the *cis*/*trans* but-2-ene interconversion, mediated by the zeolite cavity embedding the Pd atoms, is shown in Figure 4. Although the direct reaction path, occurring *via* TS3, was not found in the kinetic involving the supported Pd_2/H-ZSM-5 fragment, the mechanism, analogously to what was observed for the unsupported H-ZSM-5 system, besides the nonactivated adsorption events involves two intermediates and three transition states. Furthermore, the optimized HLBs and LLBs ground-state minima show, like in

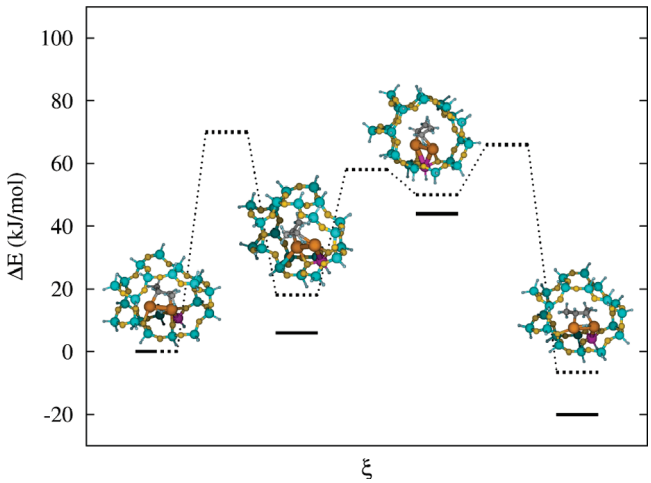


Figure 4. Catalytic pathway for the but-2-ene isomerization in the 22T Pd_2/H-ZSM-5 system and optimized ground-state HLBs minima: from the left, the reactants (*Ad-trans* and *Ad-cis*) and the intermediates (*In-trans* and *In-cis*) are shown. The *Ad-cis* species is the most stable fragment in the path: in the LLBs calculation, dotted bars, $E_0 = -8966.4881$ au, in the HLBs calculation, solid bars, $E_0 = -9009.5066$ au. Following the relative stability of the but-2-ene isomers inside the Pd_2/H-ZSM-5 fragment, the reaction coordinate, ξ, goes from the left to the right.

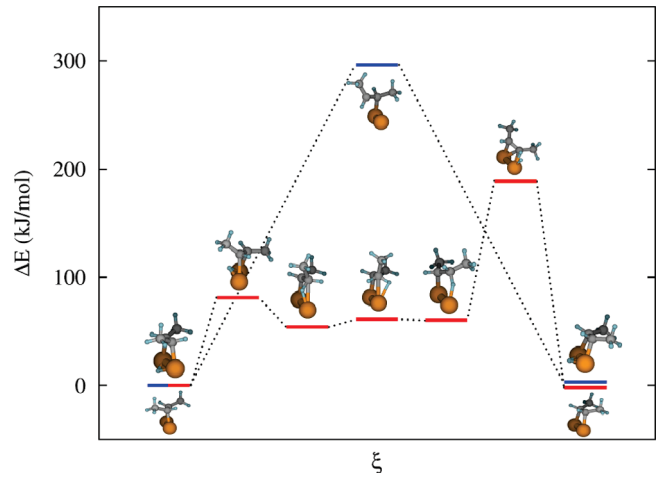


Figure 5. Pathway of the but-2-ene isomerization on bare Pd_2 (blue bars) and Pd_2H^+ (red bars) fragments. The optimized structures of the involved species are shown: down and up bars for the Pd_2 and the Pd_2H^+ processes, respectively. The *Ad-trans* and *Ad-cis* species are the most stable fragments in the Pd_2 ($E_0 = -410.7393$ au) and Pd_2H^+ ($E_0 = -411.0982$ au) reaction path. The reaction coordinate, ξ, from the left to the right follows the process *Ad-trans* → *Ad-cis*.

the unsupported zeolite system, similar geometries and comparable differential energetics (see Figure 4).

Both the *Ad-trans* and *Ad-cis* isomers are adsorbed on the two embedded Pd atoms. Consequently the double bond of the intermediates and transition states is weakened and lengthened, showing a variable extent of sp^2 to sp^3 character change. Of course, the interaction of the C2-C3 alkene group with the hydrogen, bridged between the Pd atoms in the *Ad*-species, promotes the character changes above. Also in this case the adsorbed *cis*-but-2-ene is more stable than the corresponding *trans*-but-2-ene species, by about 6 kJ/mol at the LLBs level, indicating that the reaction, inside the zeolite cavity, could be also favored in the presence of small Pd fragments.

However, the apparent activation energy concerning the formation of TS0, ca. 70 kJ/mol (see Figure 4), is considerably higher than the analogous one observed in the unsupported zeolite cavity (see Figure 3). This higher energy requirement can be rationalized considering the high affinity of the adsorbed hydrogen to the embedded palladium fragment. As a consequence, the acidic hydrogen becomes less available for the addition step, and the isomerization mechanisms inside the supported and unsupported H-ZSM-5 fragments are, to some extent, different.

The effects of the confinement and local acidity on the *cis*/*trans* bifunctional isomerization can be deepened by analyzing Figure 5, which shows the isomerization routes on the bare palladium fragments. Similar diagrams characterizing less favorable processes, not reported here for simplicity, were obtained either considering the species in triplet and quintuplet states and/or considering different but-2-ene derivative adsorption modes (namely, π and mono-σ). Clearly, the presence of one acidic center decreases the activation energies involved in the *cis*/*trans* isomerization when the same occurs on bare palladium. In particular, the protonated palladium fragment process involves equivalent adsorbed reagents, intermediates, and transition states.

Although these resemble the adsorbed hydrocarbon species observed inside the H-ZSM-5 cavity, the corresponding energetics is very different and less favorable to the occurrence of the whole process. This allows us to state that the confinement effects lead to a selective decrease of the activation energy values, mainly produced by the relative destabilization of the different derivatives adsorbed in the zeolite cavities (see also section 3.2). In the Pd_2H^+ system, the adsorbed *cis*-but-2-ene is more stable than the *trans* one, showing that besides the confinement also the local acidity could affect the stability of the adsorbed but-2-ene isomers. The role of the acidic hydrogen is also confirmed by analyzing the process occurring on the unprotonated Pd_2 fragment. In this case the mechanism changes: i) favoring the migration (2→1) of the double bond, ii) involving just one TS, showing a higher activation energy, and iii) leaving the adsorbed *trans* species more stable than the *cis* isomer.

Further considerations, being aware of the intrinsic limits of the approach, deserve the analysis of the Mulliken charges on the carbon atoms of the different but-2-ene derivatives. In Table 3 are reported the average values between topologically equivalent *cis* and *trans* carbon atoms for the different derivatives (e.g., C_4H_8) in the different definition environment (e.g., isolated), as determined by the LLBs approach. From the table, it appears that the terminal carbons (C1 and C4) are not affected by the isomerization steps. Other regular behavior is i) the increased negative charge on the C3 carbon atoms (especially in the *In* and TS species), caused by the interaction of the alkene with either the supported or the unsupported zeolite cavity surface, and ii) the values of the negative charges in the, free to rotate, -$C3H_2$-group, which in the *In* and TS species are higher than those of the fixed -$C2H_2$-group and approach the charge characterizing the terminal CH_3-group. Finally, the inversion of the charge sign

Table 3. Average Mulliken Charges of But-2-ene Species Involved in *Cis*/*Trans* Isomerization

but-2-ene derivative[a]	C1	C2[b]	C3	C4
C_4H_8	−0.35	−0.06	−0.06	−0.35
Ad(C_4H_8/H-ZSM-5)	−0.36	−0.16	−0.10	−0.38
In(C_4H_8/H-ZSM-5)	−0.33	0.13	−0.23	−0.35
TS(C_4H_8/H-ZSM-5)	−0.32	0.12	−0.24	−0.34
Ad(C_4H_8/Pd_2/H-ZSM-5)	−0.35	−0.11	−0.12	−0.33
In(C_4H_8/Pd_2/H-ZSM-5)	−0.34	−0.12	−0.21	−0.33
TS(C_4H_8/Pd_2/H-ZSM-5)	−0.34	−0.10	−0.21	−0.31

[a] In the single but-2-ene derivatives the *cis* and *trans* species show very similar Mulliken charges for carbon atoms of a given topological arrangement, irrespective of the definition environment -- i.e. isolated (C_4H_8), inside unsupported zeolite cavity (C_4H_8/H-ZSM-5), and inside supported zeolite cavity (C_4H_8/Pd_2/H-ZSM-5) -- and of the nature of the derivatives -- i.e. *Ad*, *In*, and TS. In the table are reported the average values between the *cis* and *trans* Cn, $n = 1-4$, carbon atoms for the different derivatives in the different definition environment. [b] In the C_4H_8/H-ZSM-5 systems, C2 corresponds to the carbon atom that in the TS and *In* species is bound to one O of the aluminum-modified T environment. In the C_4H_8/Pd_2/H-ZSM-5 systems, C2 is topologically not changed, with respect to the unsupported zeolite systems, after the Pd_2 addition.

in the *Ad* and TS species, involved in the isomerization inside the unsupported zeolite, has to be underlined.

Therefore, if i) the increase of the negative charge on the C2-C3 fragment is related to the sp^2 to sp^3 character change of the corresponding carbon atoms and ii) the inversion of the charge sign on the C2 atom is related to the appearing on it of carbenium character, it is possible to confirm the interpretation given, analyzing the reaction paths of Figures 3 and 4, and to state that the reaction mechanism is strictly connected to subtle changes in the confinement effects of the reaction environment and in the case of bifunctional metal-zeolite catalysts to the latter and to the mutual influence occurring between metallic centers and *BASs*.[45]

4. Conclusions

The preferential binding sites of one or more Pd atoms within a zeolite cavity are the -OH acidic groups surrounding the Al atoms. An interesting related phenomenon is represented by the reverse hydrogen spillover from the zeolite fragment and the corresponding hydrogen binding to the adsorbed Pd atoms, forming hydride species. This local transformation explains the role of the adsorbed Pd fragments on the apparent decrease of the acidic strength of the aluminum-modified zeolites and the related effects on the supported and unsupported H-ZSM-5 catalytic activity.

Concerning the *cis*-but-2-ene $\rightleftarrows$ *trans*-but-2-ene isomerization, such a reaction is very unlikely to occur in the absence of catalytic conditions, because in this case the reaction involves excited intermediate states, implicating the alkene double-bond breaking. On the other hand, the *cis*/*trans* isomerization proceeds within the H-ZSM-5 zeolite cavity because the excited-state energy required for the *in vacuo* reaction is replaced by lower ground-state activation energies, characterizing both intermediates and transition states. The *trans*-but-2-ene $\rightarrow$ *cis*-but-2-ene process could also be favored, the latter being more stable than the former inside the zeolite cavity. The findings above clearly open new perspectives in understanding catalysis of hydrocarbons with olefins as intermediates. In fact, *cis*/*trans* pathways inside zeolites have been actually hypothesized, in several catalytic reactions implying hydrocarbons, but at variance with the reported energetics here they were up till now neglected due to the involved gas phase thermodynamics.

In the zeolite catalyzed process, the *trans*- and *cis*-but-2-ene/H-ZSM-5 species present π interactions of the but-2-ene double bond with the acidic hydrogen of the zeolite. Conversely, in the structures of the intermediates and transition states, the but-2-ene double bond is reduced to a single bond. Then, one of the originally sp^2 carbons is linked to one oxygen of the aluminum-modified T environment, showing carbenium character, while the other, showing a different extent of sp^3 character, does not chemically interact with the zeolite inner surface.

The *cis*/*trans* title isomerization also occurs within the 22T H-ZSM-5 zeolite fragment in the presence of adsorbed palladium atoms. However, in this case, the intermediates and transition states show higher energy than that characterizing the corresponding species observed in the unsupported zeolite. In particular, the higher activation energy values characterizing the Pd_2/H-ZSM-5 fragment derivatives should be mainly caused by the very high affinity of the acidic hydrogen toward the same metallic centers. The suggested kinetic path on the Pd_2/H-ZSM-5 involves again two intermediates and three transition states. However, despite the fact that the *cis*-but-2-ene isomer inside the zeolite cavity embedding the palladium atoms is, as in the case of the unsupported zeolite, the most stable isomer, the isomerization mechanism in supported and unsupported zeolite appears to be different.

Acknowledgment. This work was supported by the NANOCAT Project, funded in the framework of the 6th Framework Programme of the European Community -- Contract No. NMP3-CT-2005-506621, by the University of Palermo, and by the Italian Ministero dell'Universita e della Ricerca.

Supporting Information Available: The influence of different basis sets employed to study the title reaction; the aggregation sites of the palladium atoms inside the zeolite cavity; and the multiplicity states of the supported systems. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Boudart, M.; Djega-Mariadassou, G. *Kinetics of Heterogeneous Catalytic Reactions*; Princeton U. Press: Princeton, NJ, 1984; pp 38−76. (b) Dumesic, J. A.; Rudd, D. F.; Aparicio, L. M.; Rekoske, J. E.; Trevino, A. A. *The Microkinetics of Heterogeneous Catalysis*; ACS Professional Reference Book: Washington, DC, 1993; pp 1−72. (c) Averill, B. A.; Rietjens, I. M. C. M.; van Leeuwen, P. W. N. M.; van Santen, R. A. In *Catalysis: An Integrated Approach*, 2nd ed.; van Santen, R. A., van Leeuwen, P. W. N. M., Moulijn, J. A., Averill, B. A., Eds.; Elsevier: Amsterdam, 1999; Vol. 123, Chapter 4, pp 109−206. (d) van Santen, R. A.; Neurock, M. *Molecular Heterogeneous Catalysis*; Wiley-VCH: Weinheim, 2006; pp 19−82.

(2) (a) Sauer, J. *Chem. Rev.* **1989**, *89*, 199–255. (b) van Santen, R. A.; Kramer, G. J. *Chem. Rev.* **1995**, *95*, 637–660. (c) Brändle, M.; Sauer, J. *J. Am. Chem. Soc.* **1998**, *120*, 1556–1570.

(3) Miyamoto, A.; Kobayashi, Y.; Elanany, M.; Tsuboi, H.; Koyama, M.; Endou, A.; Takaba, H.; Kubo, M.; Del Carpio, C.; Selvam, P. *Microporous Mesoporous Mater.* **2007**, *101*, 324–333.

(4) (a) Corma, A. *Chem. Rev.* **1995**, *95*, 559–614. (b) Tao, Y.; Kanoh, H.; Abrams, L.; Kaneko, K. *Chem. Rev.* **2006**, *106*, 896–910.

(5) Delahay, G.; Coq, B. In *Zeolites for Cleaner Technologies*; Guisnet, M., Gilson, J.-P., Eds.; Imperial College Press: London, 2002; Vol. 3,Chapter 16, pp 345−374.

(6) (a) Moscou, L. In *Introduction to Zeolite Science and Practice*; van Bekkum, H., Flanigen, E. M., Jansen, J. C., Eds.; Elsevier: Amsterdam, 1991; Vol. 58, Chapter 1, pp 1−12. (b) Corma, A.; Orchillés, A. V. *Microporous Mesoporous Mater.* **2000**, *35*−*36*, 21–30.

(7) (a) Damin, A.; Bonino, F.; Ricchiardi, G.; Bordiga, S.; Zecchina, A.; Lamberti, C. *J. Phys. Chem. B* **2002**, *106*, 7524–7526. (b) Boronat, M.; Concepción, P.; Corma, A.; Renz, M.; Valencia, S. *J. Catal.* **2005**, *234*, 111–118.

(8) Wang, H.; Turner, E. A.; Huang, Y. *J. Phys. Chem. B* **2006**, *110*, 8240–8249.

(9) Demuth, T.; Rozanska, X.; Benco, L.; Hafner, J.; van Santen, R. A.; Toulhoat, H. *J. Catal.* **2003**, *214*, 68–77.

(10) Nieminen, V.; Sierka, M.; Murzin, D. Y.; Sauer, J. *J. Catal.* **2005**, *231*, 393–404.

(11) Tuma, C.; Sauer, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3955–3965.

(12) (a) Kazansky, V. B.; Frash, M. V.; van Santen, R. A. *Catal. Lett.* **1997**, *48*, 61–67. (b) Kazansky, V. B. *Catal. Today* **1999**, *51*, 419–434. (c) Sommer, J.; Jost, R. *Pure Appl. Chem.* **2000**, *72*, 2309–2318. (d) Truitt, M. J.; Toporek, S. S.; Rovira-Truitt, R.; White, J. L. *J. Am. Chem. Soc.* **2006**, *128*, 1847–1852.

(13) Milas, I.; Nascimento, M. A. C. *Chem. Phys. Lett.* **2006**, *418*, 368–372.

(14) (a) Xu, T.; Barich, D. H.; Goguen, P. W.; Song, W.; Wang, Z.; Nicholas, J. B.; Haw, J. F. *J. Am. Chem. Soc.* **1998**, *120*, 4025–4026. (b) Wang, X.; Carabineiro, H.; Lemos, F.; Lemos, M. A. N. D. A.; Ramoa Ribeiro, F. *J. Mol. Catal. A: Chem.* **2004**, *216*, 131–137. (c) Boronat, M.; Viruela, P. M.; Corma, A. *J. Am. Chem. Soc.* **2004**, *126*, 3300–3309.

(15) Gleeson, D. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 579–585.

(16) Vayssilov, G. N.; Rösch, N. *Phys. Chem. Chem. Phys.* **2005**, *7*, 4019–4026.

(17) Ivanova Shor, E. A.; Nasluzov, V. A.; Shor, A. M.; Vayssilov, G. N.; Rösch, N. *J. Phys. Chem. C* **2007**, *111*, 12340–12351.

(18) (a) Teunissen, E. H.; van Santen, R. A.; Jansen, A. P.; van Duijneveldt, F. B. *J. Phys. Chem.* **1993**, *97*, 203–210. (b) Sauer, J.; Ugliengo, P.; Garrone, E.; Saunders, V. R. *Chem. Rev.* **1994**, *94*, 2095–2160. (c) Teunissen, E. H.; Jansen, A. P.; van Santen, R. A. *J. Phys. Chem.* **1995**, *99*, 1873–1879. (d) Ivanova Shor, E. A.; Shor, A. M.; Nasluzov, V. A.; Vayssilov, G. N.; Rösch, N. *J. Chem. Theory Comput.* **2005**, *1*, 459–471.

(19) Barone, G.; Casella, G.; Giuffrida, S.; Duca, D. *J. Phys. Chem. C* **2007**, *111*, 13033–13043.

(20) (a) Solans-Monfort, X.; Sodupe, M.; Branchadell, V.; Sauer, J.; Orlo, R.; Ugliengo, P. *J. Phys. Chem. B* **2005**, *109*, 3539–3545. (b) Rozanska, X. X.; Barbosa, L. A. M. M.; van Santen, R. A. *J. Phys. Chem. B* **2005**, *109*, 2203–2211. (c) Jansang, B.; Nanok, T.; Limtrakul, J. *J. Mol. Catal. A: Chem.* **2007**, *264*, 26433–26439.

(21) Milas, I.; Nascimento, M. A. C. *Chem. Phys. Lett.* **2003**, *373*, 379–384.

(22) Wei, J.; Prater, C. D. *Adv. Catal.* **1962**, *13*, 203–392.

(23) Ivanov, P.; Papp, V. *Appl. Surf. Sci.* **2001**, *179*, 234–239.

(24) Nkosi, B.; Ng, F. T. T.; Rempel, G. L. *Appl. Catal., A* **1997**, *158*, 225–241.

(25) de Ménorval, B.; Ayrault, P.; Gnep, N. S.; Guisnet, V. *Appl. Catal., A* **2006**, *304*, 1–13.

(26) (a) Cundall, R. B.; Griffiths, P. A. *Discuss. Faraday Soc.* **1963**, *36*, 111–123. (b) Cundall, R. B.; Griffiths, P. A. *J. Am. Chem. Soc.* **1963**, *85*, 1211–1212. (c) Back, M. H.; Cvetanović, R. J. *Can. J. Chem.* **1963**, *41*, 1396–1406. (d) Back, M. H.; Cvetanović, R. J. *Can. J. Chem.* **1963**, *41*, 1406–1412. (e) Golden, D. M.; Egger, K. W.; Benson, S. W. *J. Am. Chem. Soc.* **1964**, *86*, 5416–5420. (f) Egger, K. W.; Golden, D. M.; Benson, S. W. *J. Am. Chem. Soc.* **1964**, *86*, 5420–5424. (g) Benson, S. W.; Egger, K. W.; Golden, D. M. *J. Am. Chem. Soc.* **1965**, *87*, 468–476.

(27) Kotz, J. C.; Treichel, P. M.; Weaver, G. C. *Chemistry and Chemical Reactivity*, 6th ed.; Thomson Learning: Philadelphia, 2006; pp 729−731.

(28) Delbecq, F.; Zaera, F. *J. Am. Chem. Soc.* **2008**, *130*, 14924–14925.

(29) Cano, M. L.; Hamilton, D. M., Jr.; Thomason, T. B. Production of 1-Alkenes from Mixed Olefin Streams using Catalytic Distillation, U.S. Patent 7355087, 2005.

(30) van Santen, R. A. *J. Mol. Catal. A: Chem.* **1997**, *115*, 405–419.

(31) (a) Baerlocher, C.; Meyer, W. M.; Olson, D. H. *Atlas of Zeolite Framework Types*, 5th ed.; Elsevier: Amsterdam, 2001. (b) *Database of Zeolite Structures*. http://www.iza-structure.org/databases/ (accessed Aug 1, 2008). (c) *Database of Zeolite Structures, Zeolite Framework Types*. http://izasc.ethz.ch/fmi/xsl/IZA-SC/ft.xsl (accessed Nov 5, 2007).

(32) The employed notation follows the topological nomenclature of the MFI framework structure;[31] the first and the second T are coincident with the Al and the Si center, respectively.

(33) 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.; Komáromi, 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 D.02*; Gaussian, Inc.: Wallingford, CT, 2005.

(34) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (b) Stephens, P. J.; Devlin, J. F.; Chabalowsky, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(35) Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008**, *4*, 297–306.

(36) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.

(37) (a) Schreiner, P. R. *J. Am. Chem. Soc.* **1998**, *120*, 4184–4190. (b) Cramer, C. J. *J. Am. Chem. Soc.* **1998**, *120*, 6261–6269. (c) Schreiner, P. R.; Prall, M. *J. Am. Chem. Soc.* **1999**, *121*, 8615–8627. (d) Clark, A. E.; Davidson, E. R.; Zaleski, J. M. *J. Am. Chem. Soc.* **2001**, *123*, 2650–2657. (e) Graefenstein, J.; Kraka, E.; Filatov, M.; Cremer, D. *Int. J. Mol. Sci.* **2002**, *3*, 360–394.

(38) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222. (b) Francl, M. M.; Petro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, D. J.; De Frees, M. S.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(39) (a) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939–947. (b) Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. *J. Am. Chem. Soc.* **1982**, *104*, 2797–2803.

(40) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.

(41) Raghavachari, K.; Trucks, G. W. *J. Chem. Phys.* **1989**, *91*, 1062–1065.

(42) Foresman, J. B.; Frisch, Æ. *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian Inc.: Pittsburgh, PA, 1996; pp 61−90.

(43) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49–56.

(44) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(45) Kubicka, D.; Kumar, N.; Venäläinen, T.; Karhu, H.; Kubickova, I.; Österholm, H.; Murzin, D. Yu. *J. Phys. Chem. B* **2006**, *110*, 4937–4946.

(46) Yoda, E.; Kondo, J. N.; Domen, K. *J. Phys. Chem. B* **2005**, *109*, 1464–1472.

(47) Roos, B. O. In *Computational Photochemistry*; Olivucci, M., Ed.; Elsevier: Amsterdam, 2005; Vol. 16, Chapter 10, pp 317−348.

(48) Moc, J.; Musaev, D. G.; Morokuma, K. *J. Phys. Chem. A* **2008**, *112*, 5973–5983.

(49) Conner, W. C., Jr.; Falconer, J. L. *Chem. Rev.* **1995**, *95*, 759–788.

(50) Bowker, M.; Bowker, L. J.; Bennett, R. A.; Stone, P.; Ramirez-Cuesta, A. *J. Mol. Catal. A: Chem.* **2000**, *163*, 221–232.

CT800402K

JCTC Journal of Chemical Theory and Computation

Theoretical Study of the Reaction Mechanism of *Streptomyces coelicolor* Type II Dehydroquinase

L. Mattias Blomberg,*,† Martina Mangold,‡ John B. O. Mitchell,† and Jochen Blumberger*,‡

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, U.K.

Received November 7, 2008

Abstract: The reaction mechanism of a type II dehydroquinase (DHQase) from *Streptomyces coelicolor* was investigated using molecular dynamics simulation and density functional theory (DFT) calculations. DHQase catalyzes the elimination of a water molecule from dehydroquinate (DHQ), a key step in the biosynthesis of aromatic amino acids in bacteria, fungi, and plants. In the DFT calculations, 10 models, containing up to 230 atoms, were used to investigate different proposals for the reaction mechanism, suggested on the basis of crystal structures and kinetic data. Probing the flexibility of the active site, molecular dynamics simulation reveals that deprotonated Tyr28 can act as the base that catalyzes the first reaction step, the proton abstraction of the *pro-S* proton at C2 of DHQ, and formation of the enolate intermediate. The computed barrier for the first transition state (TS1), 13−15 kcal/mol, is only slightly affected by the active site model used and is in good agreement with the corresponding experimental barrier of 13.4 kcal/mol for the rate-determining step. The previously proposed enol form of the intermediate is found to be significantly higher in energy than the enolate form and is thus thermodynamically not competitive. In the second and final reaction step, protonation of the hydroxyl group at C1 by His106 followed by water elimination, there is a substantial buildup of dipole moment due to the net transfer of a proton from His106 to Tyr28. A barrier for the second transition state (TS2) that fits well with the corresponding experimental barrier could only be found if the buildup of dipole moment is at least partly compensated during the second reaction step. We speculate that this could be facilitated by regeneration of the Tyr28 anion or by proton transfer to the vicinity of His106 *before* TS2 is reached. A revised mechanism for type II DHQase is discussed in light of the results of the present calculations.

1. Introduction

The shikimate pathway is a biosynthetic route producing among others the aromatic amino acids phenylalanine, tyrosine, and tryptophan.[1−4] Chorismate is the end product of the shikimate pathway and the starting point of the synthesis of these aromatic compounds. The shikimate pathway is present in bacteria, fungi, and plants, but absent in mammals, who have to derive aromatic compounds from their diet. Hence, the shikimate pathway enzymes are attractive targets for herbicides, fungicides, and antibiotics. As an example, we mention the broad-spectrum herbicide glyphosate, an inhibitor of the penultimate step of the shikimate pathway,[5] which is, however, not suitable as an antibiotic.

The third step of the shikimate pathway, the reversible dehydration of 3-dehydroquinate (DHQ) to 3-dehydroshikimate (DHS) catalyzed by 3-dehydroquinate dehydratase, often abbreviated to dehydroquinase (DHQase, E.C. 4.2.1.10), has been an attractive target for recent inhibitor design. There are two types of DHQases, type I and II, and both catalyze

* Corresponding author e-mail: blomberg.mattias@gmail.com; jb376@cam.ac.uk.

† Unilever Centre for Molecular Science Informatics.

‡ Department of Chemistry.

10.1021/ct800480d CCC: $40.75
Published on Web 04/16/2009

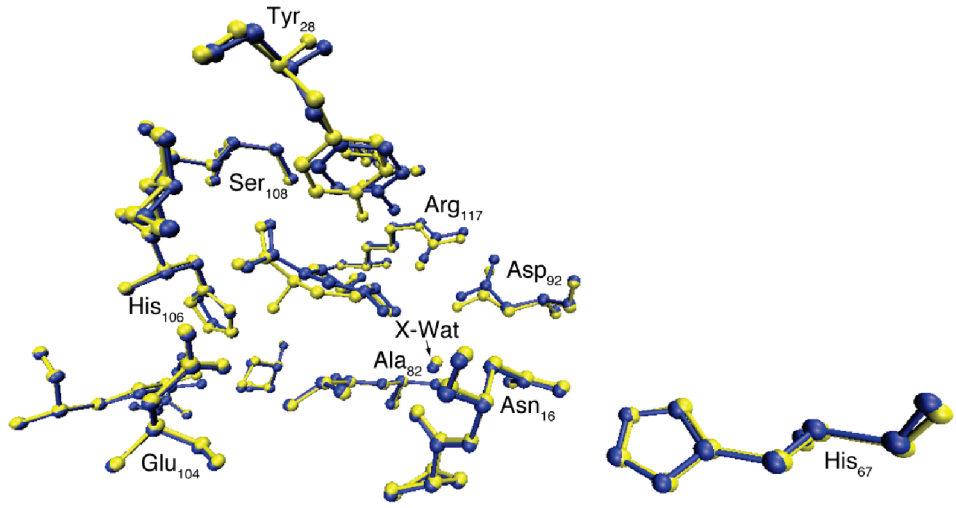


Figure 1. Aligned X-ray structures of DHQase in *S. coelicolor*. The wild-type with 2,3-anhydro-quinic acid bound is in yellow, and the R23A mutant with DHS bound is in blue.[16] The rmsd between the two structures is only 0.15 Å in the active site region.

the same overall reaction but are structurally unrelated and use different reaction mechanisms for their catalysis.[6,7] Some bacteria, such as *Escherichia coli* or *Salmonella typhi*, have type I enzymes,[8] whereas other bacteria, such as *Streptomyces coelicolor*,[9] *Mycobacterium tuberculosis*,[10] and *Helicobacter pylori*,[11] have type II enzymes. This type of convergent evolution is very uncommon for enzymes in central metabolic pathways.[12] The type II enzymes catalyze the trans dehydration of 3-dehydroquinate via a proposed enol/enolate intermediate,[13] whereas the type I enzymes catalyze a cis dehydration via an imine intermediate utilizing a conserved lysine residue.[8]

Type II DHQase is a dodecamer consisting of four trimeric units. DHQase from both *S. coelicolor* and *M. tuberculosis* can be dissociated into trimers, which were shown to be catalytically active.[14] The active site in DHQase is located in a cleft at the C-terminal end of the five-stranded parallel β-sheet. The binding pocket is covered by a loop containing residues 21−31. The loop has been shown to be flexible and closed upon binding of the substrate.[15] Furthermore, the active site of a subunit is located close to the interface of a neighboring subunit. The binding conformation of the substrate was revealed in the structures of both the wild-type enzyme forming a complex with the transition-state analogue 2,3-anhydro-quinic acid and also in an inactive R23A mutant, in which the product DHS is bound[16] (Figure 1). The substrate is held in place by about 20 hydrogen bonds. A schematic picture of the hydrogen bond network of the active site is shown in Figure 2.

While several type II DHQase inhibitors have been synthesized with affinities ranging from micromolar to nanomolar,[17−21] the type II reaction mechanism is far from being understood. The results of kinetic isotope studies are not consistent with a concerted E2 mechanism but indicate a stepwise E1cB mechanism involving an enolate intermediate.[13] This work and the analysis of crystal structures of *S. coelicolor* in complexes with a number of ligands (PDB code 1GU1)[22] have led to the proposal of a mechanism of DHQase (Figure 3). Tyr28 is assumed to have an unusually low pK_a due to electrostatic interaction with the neighboring positively charged Arg113 residue. This leads, at least to some extent, to deprotonation at physiological pH, thereby providing the base (tyrosinate) that abstracts the *pro-S* hydrogen on C2 of DHQ. The crystal structure of DHQase from *S. coelicolor* together with the measured kinetic isotope effects of DHQase from *A. nidulans* suggest that the enolate intermediate formed is converted into the enol form through proton transfer from a conserved water molecule. The latter is assumed to be regenerated by proton transfer from the amino group of the side chain of Asn16.[16] In the last step of the elimination reaction, the hydroxyl group on C1 is protonated by His106, which leads to cleavage of the C−O bond and formation of the product.

There are a number of open questions regarding the reaction mechanism described above. Not only is the protonation state of Tyr28 uncertain but also the position of Tyr28 in the crystal structure is rather unfavorable for deprotonation from the C2 atom of the substrate. Moreover, the suggested formation of an enol intermediate requires the transient formation of an anion that is likely to be high in energy (hydroxide or deprotonated side chain of an Asn residue). The formation of an enol intermediate has been invoked to explain the solvent isotope effect in DHQase from *A. nidulans*, which showed that two protons are involved in the rate-limiting step.[13] The first proton would be the one transferred from Asn16 via the crystal water to the enolate form of DHQ. The second proton would be the one transferred from His106 in the elimination step. This also implies that both reaction steps occur at similar rates, since both should be, or at least become, rate limiting for their deuterated species, respectively. However, in the same kinetic isotope study for DHQase from *M. tuberculosis*, the corresponding solvent isotope effect indicates the involvement of only one proton in the rate-limiting step.[13] Thus, there is a discrepancy in the mechanisms comparing the two different species.

In the present theoretical study, we investigate the energetic feasibility of various variants of the E1cB reaction

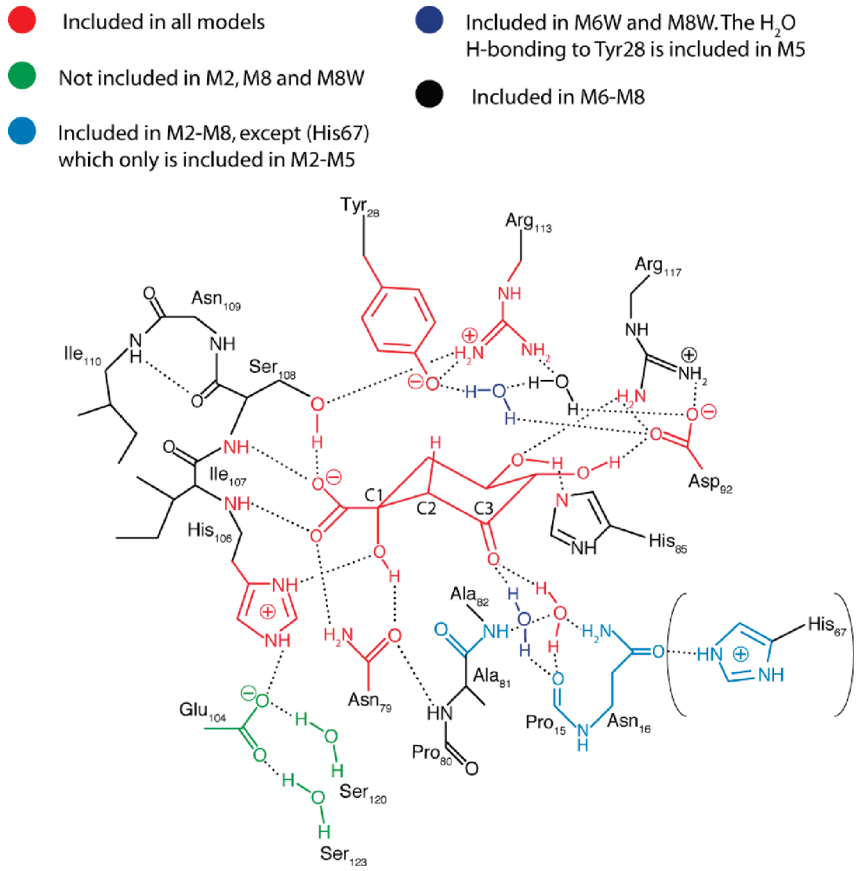


Figure 2. Schematic picture of the active site in DHQase binding the substrate DHQ. The differences in the models used are color coded.

mechanism described above using QM calculations complemented with dynamical information obtained from molecular dynamics simulation of the substrate bound complex. Preliminary investigations on the E2 mechanism showed that the barrier for breaking the C−H and C−OH bonds simultaneously is unfavorable by about 10 kcal/mol, as compared to the E1cB type of mechanism for the corresponding model. This is in line with experimental evidence suggesting that the E2 mechanism does not play an important role.[13] Carrying out long molecular dynamics runs, we find that at room temperature rare thermal fluctuations occasionally bring Tyr28 within hydrogen-bonding distance of the proton at the C2 atom of the substrate. Thus, if Tyr28 is deprotonated, it can function as the base that catalyzes the first reaction step in the E1cB mechanism. The QM calculations show that the deprotonation occurs directly; the water-assisted deprotonation is disfavored by 10 kcal/mol. Moreover, the enol form proposed for the intermediate is found to be too high in energy to compete with the enolate form. The barriers for the reaction steps are obtained in reasonably good agreement with experiment, even though they are rather sensitive to the size of the model used. A revised picture of the reaction mechanism based on the calculations presented herein is given at the end of this article.

2. Methods

2.1. Molecular Dynamics Simulation. We chose a trimeric subunit of DHQase as a model for molecular dynamics simulation. The trimer structure was obtained from the X-ray crystal structure of dodecameric DHQase from (PDB code 1GU1)[22] by selecting chains A, B, and C and the crystal water molecules assigned to them. All three chains were treated as identical. The transition-state analogue 2,3-anhydroquinic acid (AHQ) bound to the active site was modified to DHQ by addition of a carbonyl oxygen atom at C3 and a hydrogen atom at C2. In the crystal structure, two buffer molecules bound in close proximity to the active site (one glycerol molecule and tartaric acid, respectively). They were removed and replaced by nine water molecules. The water molecules were placed so that the oxygen atoms were on top of the crystallographic positions of the oxygen atoms of the two buffer molecules. Two positions for the side chains of Ser30, Glu124, and Glu143 were reported; conformation "A" was taken in this work. In the crystal structure, residues 151−156 were missing; Gly150 was taken as the C-terminus. Hydrogen atoms were added using the program "xleap" included in the AMBER9[37] simulation package. Protonation states of ionizable side chains were chosen corresponding to pH = 7. Tyr28 was modeled in its deprotonated form, histidines 47, 67, 73, and 118 were protonated at the εN atom, histidines 58, 85, 111, and 119 were protonated at the δN atom, and His106 was doubly protonated. The protonation states of the histidine residues were chosen after inspection of the local hydrogen bond network and the extent of exposure to the solvent. The protein was neutralized by adding 21 K^+ ions and solvated by adding 14 753 water molecules and 33 K^+ and Cl^- ions. The molality of KCl was 0.10 mol/kg; the total number of atoms was 52 845.

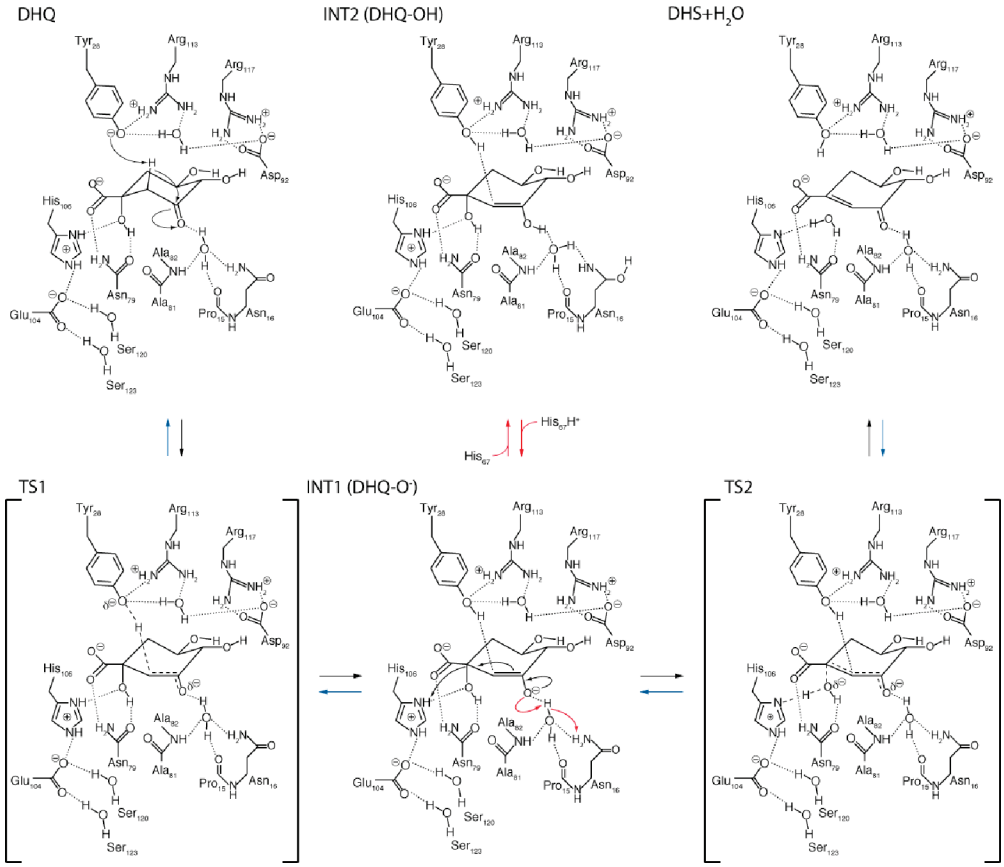


Figure 3. Schematic picture of the proposed reaction mechanism catalyzed by type II DHQase.

The MD simulations were carried out using the AMBER99 force field and the TIP3P model for water. A set of point charges reproducing the electrostatic potential for DHQ and deprotonated Tyr28 was obtained from DFT calculations at the B3LYP/cc-PVTZ(-f)//6-31G(d) IEFPCM (r = 1.4 Å, ε = 4) level of theory. Atomic point charges representing the electrostatic potential from the DFT calculation were obtained according to the RESP method. Force constants for bonds and angles in DHQ were extracted from the Hessian calculated at the B3LYP/6-31G(d)[25,26] level of theory. Seminario's algorithm, which is fully invariant to the choice of internal coordinates,[23] was used to extract force constants for bonds and angles in the optimized structure. The dihedral force constants were taken from the GAFF force field implemented in AMBER.

The atomic positions of DHQ and the nine water molecules replacing the two buffer molecules were relaxed by energy minimization while keeping all protein atoms fixed at their crystallographic positions. All solvent molecules were equilibrated for 50 ps in the NPT ensemble using temperature rescaling to 300 K and a target pressure of 1 bar. Then the protein was equilibrated first by carrying out four consecutive runs of length 100 ps/run using harmonic restraints on the protein atom positions with force constants of 99, 50, 25, and 10 kcal/(mol Å^2). The restraining potentials were removed, and the free protein was equilibrated for 1 ns in the NPT ensemble using a Berendsen thermostat with target temperature 300 K. During this final equilibration period and the subsequent production runs, the center of mass of each of the three subunits was restrained to the value of the crystal structure. The center of mass restraint was applied to avoid tumbling motion of the trimer that would have brought protein atoms too close to the edge of the simulation cell. Using the restraint the smallest distance between any protein atom and the box edge was always larger than 10 Å as compared to the cutoff of 12 Å used for truncation of short-range nonbonded interactions. Production runs used for calculation of thermal averages were carried out for 5 ns in the NVT ensemble using a MD time step of 2 fs. The cell dimensions were 94.03 × 90.34 × 62.69 Å^3. The r-RESPA multiple time stepping method was used with bonded and short-range nonbonded forces calculated every time step and electrostatic forces every second time step. The SPME algorithm[24] with default parameters was used for the latter. All molecular dynamics simulations were carried out with the NAMD[38] package.

2.2. QM Calculations. In the present study, 10 QM models were investigated, each based on the crystal structure of *S. coelicolor*, PDB code 1GU1. The models were chosen so as to investigate different features of the active site and so as to test different propositions regarding the reaction mechanism. The active site in DHQase is large and contained a complex hydrogen-bonding network due to the carboxylate, carbonyl, and the three hydroxyl groups in DHQ (Figure 2). Models M1−M5 focus on modeling the chemically important parts of the active site as accurately as possible, whereas interactions that do not change significantly during the reaction were modeled more approximately. Models M6−M8 are computationally more demanding and incorporate a larger part of the surroundings of the active site. A summary of

Table 1. Differences among Models M1−M8; See Also Figure 2[a]

	M1	M2	M3	M4	M5	M6	M6W	M7	M8	M8W
His67	NM	P	P	NP	P	NM	NM	NM	NM	NM
Glu104	*	NM	*	*	*	*	*	P	NM	NM
Ser120	*	NM	*	*	*	*	*	*	NM	NM
Ser123	*	NM	*	*	*	*	*	*	NM	NM
H_2O[b]	NM	NM	NM	NM	NM	*	*	*	*	*
2 × H_2O[c]	NM	NM	NM	NM	*[d]	NM	*	NM	NM	*

[a] NM - not modeled; * - modeled; P - protonated; NP - not protonated. [b] Water molecule included in models M6−M8, which is hydrogen bonding to $TyrO^-$. [c] Two additional water molecules added in M6W and M8W. One is hydrogen bonding to $TyrO^-$, and the other is hydrogen bonding to the carbonyl oxygen on C3 in DHQ. [d] Note that in M5 only one water is included, which is involved in the proton transfer between DHQ and $TyrO^-$.

the models used is given in Table 1 and Figure 2. Furthermore, a detailed description of the models is given in the Appendix. For all models, the QM region was terminated by saturating the bond that is cut with a hydrogen atom. The latter was positioned along the axis of the bond that is cut at a typical X−H bond length. The reaction energy profile for all model systems was computed using reaction coordinate driving combined with geometry optimization. A small model system (M0) was used for calculation of thermal and entropic contributions to the reaction energies.

The geometries of all the stationary points in the reaction studied were optimized using Becke's[25] three-parameter hybrid exchange functional combined with the Lee−Yang−Parr[26] correlation functional (B3LYP). The calculations were performed using the Gaussian03[27] suite of programs. Geometries and Hessians were obtained using Pople's split valence basis set 6-31G*. Models M1−M5 were optimized taking the dielectric effect of the protein into account by using the integral equation formalism model (IEFPCM) implemented in Gaussian03.[28−30] The dielectric constant was set equal to 4.[31] Bondi's atomic radii were used together with a probe radius of the solvent of 1.4 Å to create a cavity around the solute. Because of their size (∼225 atoms), models M6−M8 were optimized in the gas phase and dielectric effects were included only for the final single point calculations. The final energies for the optimized geometries were evaluated using Pople's triple split valence basis set 6-311+G(2d,2p), containing double polarization on first- and second-row atoms and one diffuse function on second-row atoms. To keep the main features of the structure close to the crystal structure, restrictions on certain atoms were applied during geometry optimization. Either the position of C_α was fixed if included in the model, or the position of the terminating hydrogen atom was fixed. The small rmsd (0.15 Å) between the crystal structure with the transition-state analogue bound and the crystal structure with the product bound (Figure 1) indicates that this restriction on the flexibility of the active site is an adequate approximation for this enzyme. The smallest model (M0) was used for computation of thermal and entropic contributions of the free energy profile. The same corrections were used for all models. The accuracy of the B3LYP functional was tested in the extended G3 benchmark set,[32] which consists of enthalpies of formation, ionization potentials, electron affinities, and proton affinities for molecules containing first- and second-row atoms. In these tests, the B3LYP functional gave an average absolute error of 4.3 kcal/mol for 376 different entries.[32] The calculated enthalpy for the reaction DHQ → DHS + H_2O in aqueous solution was −1.7 kcal/mol, to be compared with the experimental value of 0.6 kcal/mol.[33]

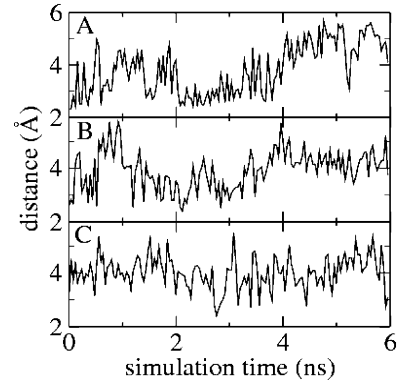


Figure 4. Distance between the oxygen atom of deprotonated Tyr28 and the *pro-S* hydrogen atom at C2 of the substrate DHQ as obtained from molecular dynamics simulation at 300 K. Panels A, B, and C show the distance for each of the three equivalent subunits of the DHQase trimer simulated.

3. Results

3.1. MD Simulation of Substrate-Bound DHQase (DHQ (E)). MD simulations of DHQase were carried out to investigate the structure and dynamics of the hydrogen bond network of the substrate bound state. The simulation of trimeric DHQase gives a stable protein structure as indicated by the small rmsd fluctuating between 1.5 and 1.8 Å relative to the crystallographic position of the trimer subunit in dodecameric DHQase. This shows that the simulation of an isolated trimer subunit does not cause any artificial conformational changes, thereby justifying our approach. The substrate DHQ is held tightly in the active site by about 20 hydrogen bonds formed with Ala81, Ala82, Arg117, Asn79, His85, His106, Ile107, Ser108, Tyr28, one crystallographic water molecule, and one to two further solvent molecules.

In the following, we discuss three regions of interest in more detail: First, the solvation structure of the deprotonated tyrosinate Tyr28; second, the solvation structure of the carbonyl oxygen atom of DHQ; and finally, the hydrogen-bonding network of the leaving group C1−OH. The tyrosinate anion is stabilized by formation of two strong hydrogen bonds with the guanidinium group of Arg113. The mean O···H hydrogen bond lengths are 2.06 and 2.16 Å, respectively. Additional stabilization is gained by hydrogen bond formation with solvent molecules. The water molecules in the channel connecting Tyr28 with the bulk solvent are very mobile, and the water molecule binding to Tyr28 is frequently replaced by a neighboring water molecule. The average distance between the tyrosinate oxygen atom and the proton at C2 is large, 3.81 (±) 0.94 Å, see Figure 4. The smallest distance observed for several configurations along the 6-ns trajectory was 2.3 Å. These events, which are coupled with the dynamics of hydrogen bond break/formation

Table 2. Free Energies Given in Kilocalories per Mole for the 10 QM Models Used in the Present Study[a,b]

	ΔΔG[a]	M1	M2	M3	M4	M5	M6	M6W	M7	M8	M8W
DHQ(aq)	-	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1
DHQ(E)	0	0	0	0	0	0	0	0	0	0	0
TS1	-1.8	14.0	13.5	15.3	14.7	28.3	14.9	14.9	18.9	17.8	12.7
DHQ-O$^-$	0.3	12.3	12.2	14.6	14.6	11.3	13.4	13.5	14.5	13.2	8.5
TS2	-1.3	17.4	14.4	20.5	20.2	-	18.7	22.5	17.8	15.7	16.1
DHS(E)	-3.9	0.7	-4.2	9.2	8.0	-	-0.2	4.6	-5.9	-8.2	-3.3
DHS(aq)	-	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2

[a] Gibbs free energy correction, which was calculated for a small unrestricted model of the DHQase active site and used in all models. [b] In the columns colored light green the charged Glu104 moiety is included, whereas in the columns colored light yellow the moiety is excluded or the charge is cancelled (M7). In the columns colored white a charged His67 is included in addition to the Glu104 moiety. The binding energy and the free energy of the reaction in aqueous solution are calculated from experimental equilibrium constants.[33,36]

between the oxyanion and solvent molecules, could eventually lead to deprotonation of the substrate. Thus, our simulations show that Tyr28 has sufficient conformational flexibility to overcome the relatively large separation distance between the deprotonated side chain oxygen and the C2-proton of the substrate (TyrO···C2 distance = 3.35 Å in the crystal structure 1GU1).

The carbonyl oxygen atom of DHQ forms on average two hydrogen bonds, one with a conserved crystal water and one with a solvent molecule. As was the case for Tyr28, the solvent molecule is very mobile and frequently gets replaced by another solvent molecule. The crystal water is tightly bound in the enzymatic pocket over the entire 6 ns of the present simulation. The side chain of Asn16 and the backbone atoms of Pro15 and Ala82 hold the crystal water in a rigid position. The hydrogen bond contacts between the substrate, crystal water, and the side chain of Asn16 could allow for the proposed proton transfer from Asn16 to the carbonyl oxygen atom after deprotonation of the substrate.

The hydrogen atom of the leaving group C1−OH forms a strong hydrogen bond with the carbonyl oxygen atom of Asn79 throughout the simulation. This ensures a correct orientation of the leaving group that eventually facilitates proton transfer from the Nδ of His106 to the oxygen atom of C1−OH. Thus, our molecular dynamics simulations show that the hydrogen bond network of substrate-bound DHQase is stable against thermal fluctuations. The general base Tyr28 and general acid His106 are connected to the bulk solvent via a chain of highly mobile water molecules that facilitate the shuffling of protons to and from the active site.

3.2. QM Calculations. The free energy profiles obtained from QM calculations of model systems M1−M8 are summarized in Table 2 and Figures 5 and 6 and are presented and discussed in the following sections.

Formation of Intermediate DHQ (E) → DHQ−Enolate. In the optimized structures of DHQ bound in the active site (M1−M8), the Tyr28 oxygen is located at a distance of 2.48−3.45 Å from the *pro-S* hydrogen of C2, and 3.39−4.39 Å from C2, in DHQ, respectively. The distance in the crystal structure corresponding to the latter is 3.35 Å, close to the lower values in the optimized range. The shorter distances were obtained for the larger models M6−M8, and especially in the models incorporating two additional waters (M6W and M8W) the distance approaches the experimental results. The transition state for proton transfer from C2 to the tyrosinate

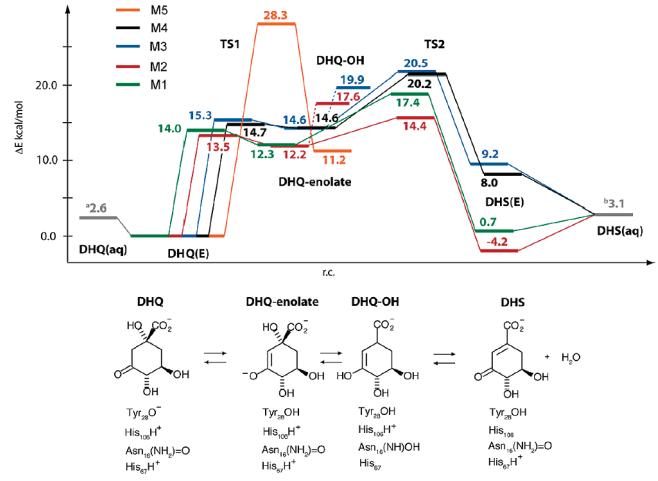


Figure 5. Computed energy profile for models M1−M5 for the reaction mechanism of type II DHQase. See Figure 2 for the models used. The apparent binding energy of DHQ in DHQase is $\Delta H_{app} = -2.6$ kcal/mol ($K_{app} = 5.40$ mM).[35] The standard enthalpy change of the overall reaction is $\Delta_r H^\circ_m = 0.55 \pm 0.55$ ($K = 4.6 \pm 1.5$).[32]

oxygen (TS1) is reached late, at an O−H bond distance of about 1.15 Å. The C2−C3 bond is shortened from 1.54 to 1.43 Å in TS1, indicating partial formation of a double bond. Very similar values for the C2−H distance (about 1.53 Å) are obtained for models M1−M8. Interestingly, the structure of TS1 is similar in all models even though the models differ substantially and M6−M8 are optimized in the gas phase.

The mean value of the free energy barrier for the proton transfer is 15.2 ± 2.0 kcal/mol, spanning values between 12.7 and 18.9 kcal/mol as obtained for M8W and M7, respectively. Note that the energies for M5 are excluded, since this model corresponds to the unfavorable assisted deprotonation mechanism discussed further below. If the models with the largest barriers are excluded (M7 and M8), the mean barrier is slightly smaller and the uncertainty is significantly reduced, 14.3 ± 0.9 kcal/mol. This value is in good agreement with the experimental rate-determining barrier of 13.4 kcal/mol, calculated from the turnover rate of 960 s^{-1} in wild-type *S. coelicolor* using transition-state theory. The barriers for M7 and M8 are too large because there is only the crystal water stabilizing the buildup of negative charge on the carbonyl oxygen atom. The second water molecule that, according to MD simulations, binds to

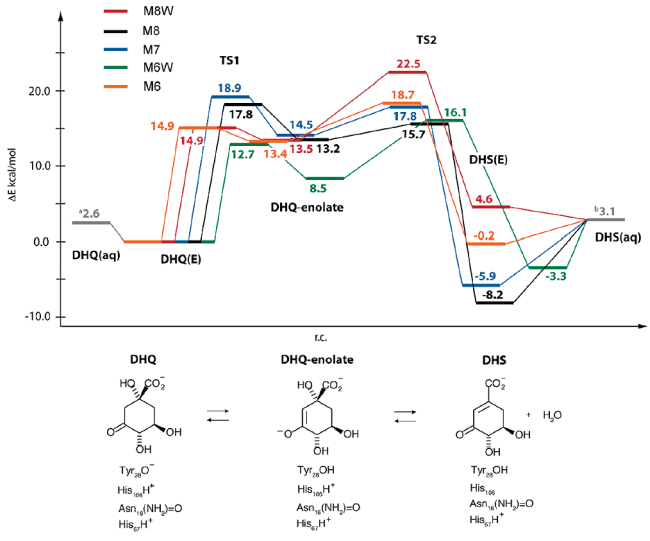


Figure 6. Computed energy profiles for models M6−M8 for the reaction mechanism of type II DHQase. See Figure 2 for the models used. The apparent binding energy of DHQ in DHQase is $\Delta H_{app} = -2.6$ kcal/mol ($K_{app} = 5.40$ mM).[36] The standard enthalpy change of the overall reaction is $\Delta_r H°_m = 0.55 \pm 0.55$ ($K = 4.6 \pm 1.5$).[33]

the carbonyl oxygen atom in the ES complex was not modeled. Indeed, inclusion of this second water molecule in model M8W reduces the barrier to a value consistent with the experimental rate. The smaller models M1−M4 give similarly good values as a consequence of two opposing effects. Both the water stabilizing deprotonated Tyr28 and the water stabilizing the buildup of negative charge in the transition state are not modeled. The effect is that the energies of reactant and transition state are raised by approximately equal amounts.

The formation of the enolate intermediate (DHQ−enolate) is endergonic with a mean value of 12.8 ± 1.9 kcal/mol, spanning from 8.5 to 14.6 kcal/mol for M8W and M3/M4, respectively. The deviation of model M8W from the mean value of the enolate intermediate is larger than for other models, and excluding M8W gives a mean value of 13.3 ± 1.2 kcal/mol. Thus, with the exception of M8W, the endergonicity of forming the enolate intermediate is less affected by the choice of the model then the preceding transition state. The endergonicity of this reaction step is the origin of the late transition state in TS1. The enolate intermediate is high in energy compared to $TyrO^-$, since there is no stabilization by a positively charged residue, such as Arg113 in the case of Tyr28. In M8W, the two extra water molecules modeled not only lower the barrier but also increase the stability of the enolate intermediate. During the reaction, the C3−O bond length of the carbonyl group is slightly increased, from 1.22 Å in DHQ to 1.25−1.30 Å in DHQ−enolate, whereas the C2−C3 bond is shortened from 1.51−1.52 to 1.38−1.44 Å. Interestingly, protonated Tyr28 remains hydrogen-bonded to the C2 atom of DHQ−enolate at a distance of 1.90 Å in models M1−M8, indicating that a significant part of the negative charge density remains on the C2 in the enolate intermediate. The corresponding distance in models M6W and M8W is 2.04−2.08 Å, indicating that more negative charge has been transferred to the oxygen atom in the models that contain the additional solvent molecule.

Isomerization of Intermediate DHQ−Enolate → DHQ−OH. As discussed above, the presence of a water molecule in the crystal structures together with solvent isotope effects indicate that two protons are involved in the rate-limiting step. This has led to the suggestion that an enol intermediate is formed instead of enolate. It has been suggested that the crystal water assists the proton transfer from Asn16 to the enolate, leaving Asn16 deprotonated.[16] However, such an intermediate was too high in energy to be plausible. Therefore, His67, which belongs to a neighboring subunit of the dodecamer structure of DHQase, was included in models M2−M5. A histidine residue is likely to be protonated at low cost at physiological pH and would thus be the most reasonable source of the proton needed to form the suggested enol intermediate. However, the formation of the enol through proton transfer from Asn16, which in turn is reprotonated by His67 to the isomeric form CNH(OH) of Asn16, is unfavorable according to our calculations. This enol intermediate is 5.4 and 5.3 kcal/mol higher in energy than the corresponding enolate intermediate in models M2 and M3, respectively. Note that the transition state leading to this state would be even higher in energy. Thus, the suggested enol intermediate does not seem likely to be of importance for the conversion of DHQ to DHS.

Formation of Product DHQ−Enolate → DHS (E) + H_2O. The second reaction step consists of protonation of the hydroxyl group at C1 by His106 and heterolytic cleavage of the C1−OH bond. This leads to the formation of the products, DHS and water. In the transition state TS2 the C1−OH bond distance is 1.67 Å for the smaller models M1−M5, which were optimized in a dielectric continuum. A significantly larger value of 1.95 Å was obtained for models M6−M8, which were optimized in the gas phase. Furthermore, the N−H bond of His106 increases from 1.06 Å in the enolate to 1.45 Å in TS2 for models M1−M5, and from 1.05−1.08 to 1.25 Å in models M6−M8. At the same time, the distance between the proton being transferred from His106 to the oxygen atom decreases from 1.62−1.63 to 1.09−1.10 Å in M1−M4 and from 1.54−1.69 to 1.19−1.27 Å in M6−M8. Tyr28 remains also in TS2 hydrogen bonded to C2. However, the hydrogen bond breaks upon formation of DHS.

The gas-phase-optimized structures are more affected by the presence of the charge on Glu104. The later transition states in the gas-phase models are caused by charge-transfer effects. After proton transfer from His106 to the DHQ−enolate, the His106-Glu104 motif is negatively charged, whereas the Tyr28-Arg113 motif is positively charged. The buildup of this net dipole is clearly disfavored in the gas phase when compared to a dielectric medium. Thus, the gas-phase models are optimized on a more endothermic energy surface, which leads to a later transition state consistent with Hammond's Postulate.[34,35] In addition to being an earlier transition state in the structure optimized in the dielectricum (M1−M4), the charge transfer is also closer to completion. The effect of the treatment of the surroundings on the reaction coordinates may seem to be large. However, the energy surface at the

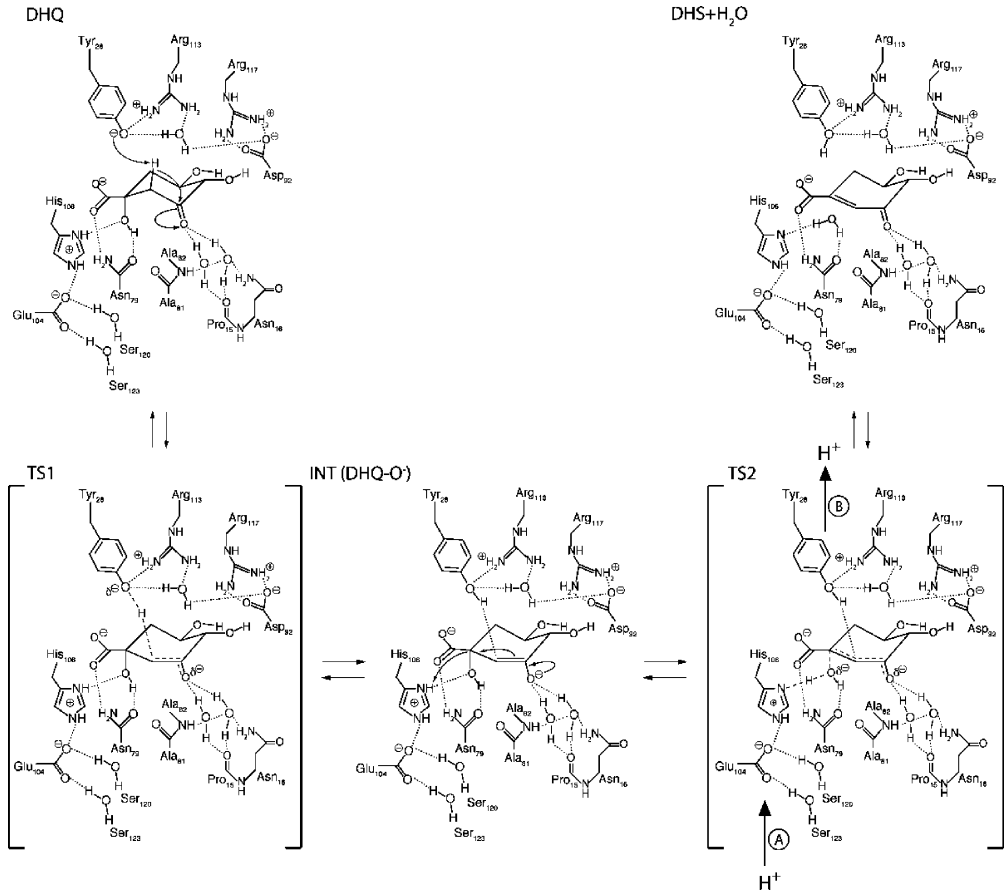


Figure 7. Revised mechanism for the dehydration of DHQ to DHS and water in DHQase. A and B indicate possible proton motions occurring before or connected to the transition state (TS2).

transition state TS2 is rather flat, and the effect on the barriers after the solvent effect has been added is rather small, as can be seen when comparing the energies for TS2 for the corresponding models (e.g., M4 and M6), displayed in Table 2. The barrier of TS2 is 5.0 ± 2.2 kcal/mol compared to that of the enolate intermediate and 18.1 ± 2.6 kcal/mol compared to that of DHQ, which is the resting state. The barrier for the different models ranges from 14.4 kcal/mol in M2 to 22.5 kcal/mol in M6W. This barrier should be compared to the experimental turnover rate of 960 s^{-1} in wild-type *S. coelicolor* corresponding to a rate-determining barrier of 13.4 kcal/mol.

The barrier for TS2 for most of the models is too high compared to that of experiment. However, as can be seen for models M2, M8, and M8W, the exclusion of Glu104 from the models leads to a reduced barrier height since a large part of the effect of the charge transfer and net dipole formation is removed. Thus, to reach a barrier height fitting well with experiment, a charge cancelation of Glu104 seems necessary. This could be realized by proton transfer through the hydrophilic channel connecting bulk solvent with Glu104. A similar effect would be obtained if Tyr28 were deprotonated at this stage and the catalytically active form of the active site regenerated. M7 has a somewhat higher barrier despite the protonated and thus uncharged Glu104. However, for this model larger structural changes occur, breaking the hydrogen bond between His106 and Glu104, causing the slightly higher barrier compared to that of M8.

The formation of the final products, enzyme-bound DHS and H_2O (DHS(E)), is exergonic by −13.7 ± 5.8 kcal/mol compared to the enolate intermediate and by 0.1 ± 6.1 kcal/mol compared to the enzyme−substrate complex DHQ(E), respectively. The overall reaction ranges from endergonic by 9.2 kcal/mol (M3) to exergonic by 8.2 kcal/mol (M8) (Table 2). This broad range of energy values is partly caused by the difference in charge-transfer effect due to the presence or absence of charged Glu104. A second cause is the difference in flexibility in the models. The structural changes in the last reaction step are large compared to those of previous steps. The smaller models, M1−M4, are much more restricted compared to models M6−M8 and will thus perform worse for the last step compared to the preceding stationary points. According to the experimental equilibrium constants for the reaction in aqueous solution and for substrate binding, unbound solvated DHS is 2.2 kcal/mol higher in energy compared to DHQ bound to the active site.

Kinetic Isotope Effects. As discussed above, solvent isotope studies imply that a different number of protons are involved in the rate-determining step for DHQases in different species. In *A. nidulans*, two protons take part in the rate-determining step, but only one in *M. tuberculosis*. One suggestion, which was shown to be energetically unfavored in the present study, involves the crystal water hydrogen bonding to the carbonyl group on C3 in DHQ. Another possibility could be that deprotonation of the substrate by Tyr28 does not occur directly but is assisted by a water molecule. In some species,

such as *A. nidulans*, water-assisted proton abstraction could occur, whereas in species such as *M. tuberculosis* no water molecule is involved. We investigated this possibility by adding an extra water molecule in model M5 and recalculating the energetics of the first reaction step. As can be seen in Table 2 and Figure 6, the extra water stabilizes the enolate intermediate, but the barrier for TS1 increased to 28.3 kcal/mol. Thus, the water-assisted proton transfer is unfavorable by 13.6 kcal/mol compared to direct proton transfer. This number is too large to explain the observed differences in solvent isotope effects for the different species. On a more speculative note, the solvent isotope effect could also be related to the proton transfer in or out of the active site before TS2 is reached. Further investigation is clearly needed to understand the kinetic differences among DHQases from different species.

4. Conclusions

The reaction mechanism of DHQase that results from our present theoretical investigation is illustrated in Figure 7. We find that deprotonated Tyr28 has enough conformational flexibility to overcome the large separation from the substrate in the crystal structure. The computed barrier for deprotonation of the substrate is in the range 13−15 kcal/mol, close to the experimental rate-limiting step corresponding to 13.4 kcal/mol. The deprotonation reaction follows a direct mechanism. The alternative, water-assisted, mechanism was found to be more than 10 kcal/mol higher in energy.

For the second reaction step, protonation of the enolate intermediate by His106 and water elimination, we find that a cancelation of the charge on Glu104 seems to be crucial for the reaction to proceed at a feasible rate. This could be accomplished by a proton entering the vicinity of Glu104, as an intermediate step in the regeneration of a protonated His106. A similar effect would probably be seen if the proton on Tyr28 is removed at this stage regenerating the catalytic tyrosinate. One could also imagine that this is coupled to the proposed motion of the loop covering the binding pocket, which may have to open upon product release.[15] The cancelation of charge on Glu104 before TS2, discussed above, would correspond to the first reaction step occurring in an environment best described by model M6W, while for the second reaction step the surroundings are best described by model M8W.

The mechanism shown in Figure 7 is, to a certain extent, similar to a previously proposed mechanism based on crystal structure analysis and kinetic data.[16] However, one important difference concerns the nature of the intermediate. The formation of an enol form of the intermediate was suggested in ref 16 to explain the solvent isotope effects of the enzyme-catalyzed reaction. Our calculations indicate, however, that the enol form is significantly higher in energy than the corresponding enolate form, showing that the enol intermediate should not be of importance for the reaction studied.

Our suggestion that shuffling of protons is needed to pass TS2 could also have implications on the interpretation of the solvent isotope effects. If the proton motions and the breaking of the C1−O bond in TS2 are tightly coupled, the proton motions may be affecting the solvent isotope effects. Different species may exhibit slightly different energetics for the proton motions, which could give rise to the differences observed for the number of protons involved in the rate-limiting step. This would also suggest that the second reaction step is rate-limiting.

Previous and present proposals for the reaction mechanism of type II DHQase assume that Tyr28 is deprotonated. If the reaction is initiated by deprotonation of the C2 atom, then there is indeed little doubt that tyrosinate is the catalytic base. This is simply because there are no other basic residues in the vicinity of the C2 proton. In future investigations, we aim at computing the pK_a value of this residue using a novel QM/MM technique. Moreover, we would like to compute the acidity constants of a number of other catalytic residues along the reaction path to corroborate and extend our present findings.

Acknowledgment. We thank the Wenner-Gren Foundations and Swedish Research Council for postdoctoral funding (L.M.B.), Unilever for funding (J.B.O.M.), the European Commission for an EST PhD studentship under the Marie Curie Actions Program (M.M.), and the Royal Society for a University Research Fellowship (J.B.). We thank the U.K.'s EPSRC National Service for Computational Chemistry Software for providing us with both access to software and computer time. We also thank Prof. Chris Abell for valuable discussions on type II DHQase.

Appendix

Model M0. The M0 model for the first reaction step (formation of enolate of DHQ) includes DHQ with a protonated carboxyl group (mimicking the stabilization given by the carboxyl binding pocket), Tyr28 modeled as phenolate, and a water molecule hydrogen bonding to the carbonyl oxygen atom of DHQ. In the second reaction step (formation of DHS and water), the phenolate molecule was removed and His106, modeled as protonated imidazole, was included.

Models M1−M5. The active site models are illustrated in Figure 2. The carboxylate binding pocket, which consists of Ile107 and Ser108, is modeled by three water molecules, mimicking the hydrogen bonds from the backbone peptides and OH side chain of Ser108. Furthermore, His85, which accepts a hydrogen bond from the OH group on C5 in DHQ, is also modeled as water molecules. The salt bridge between Arg117 and Asp92 is modeled as a formate ion and an ammonium ion, respectively. Tyr28 was cut between C_α and C_β, and Arg113 was cut between C_γ and C_δ. A crystal water molecule, which was observed in several X-ray structures of DHQase, was proposed to assist a proton transfer from Asn16 to the enolate form of DHQ, forming the corresponding enol intermediate. Because of the suggested importance of this intermediate, a larger part of the protein around the crystal water was included in the model. This water molecule can form three hydrogen-bonding contacts to surrounding residues. Hydrogen bonds can be formed to the amide group of Asn16, the backbone carbonyl of Pro15, and the backbone amide of Ala82. The backbone amide of Ala82 was modeled as formamide, and the carbonyl group of Pro15 as an acetate group forming a peptide bond with full side chain model of

Asn16. Furthermore, His67 was included in the model since the proposed deprotonation of Asn16 was shown to be energetically unfavorable. Furthermore, the proximity of His67 together with a plausible pK_a of around 7 make the latter a reasonable candidate for the presence of an extra proton in the system, which could stabilize a DHQ enol intermediate. His67 is modeled as imidazole. His106 is modeled as a protonated imidazole, and Glu104 is modeled as a formate ion. Ser120 and Ser123, which both are hydrogen bonding to Glu104, are modeled as water molecules. Asn79 is modeled as formamide.

The smallest model M1 differs from M3 by the exclusion of all residues that form hydrogen bonds to the crystal water, which is hydrogen bonding to the carbonyl oxygen atom of DHQ. In model M2, the Glu104-Ser120-Ser123 part of the active site is excluded. In model M3 His67 is protonated, whereas in model M4 it is deprotonated. In model M5, an additional water molecule is included, bridging Tyr28 and the carbonyl oxygen on C3, and later taking part in the proton transfer from DHQ to Tyr28.

Models M6−M8. For models M6−M8, an attempt was made to include all of the protein in the vicinity of the substrate. Model M6W, which consists of about 230 atoms, is shown schematically in Figure 2. The N-terminal end of peptide chain His106-Ile107-Ser108-Asn109-Ile110 was cut at the C_α−N bond between His106 and Val105. Similarly, the C_α−C(=O) bond between Ile110 and His111 at the C-terminal end was cut. Furthermore, only the backbone of Asn109 was included, and the side chain, which is pointing away from the active site, was excluded. In the peptide chain discussed above, every second C_α was restricted in the optimization with the coordinates taken from the X-ray structure (1GU1), starting with C_α in His106. The full side chain including the C_α was included in the model of Tyr28, whereas the C_α−N and C_α−C(=O) bonds to Ile27 and Gly29, respectively, were cut. Arg113 was cut between C_β and C_γ. The model of Arg117 is equivalent to the model of Arg113. Asp92 was cut between C_α and C_β. Pro15 and Asn16 were modeled as the full side chain of Asn16 and include the peptide bond to Pro15, whereas the C(=O)−C_α in Pro15 and C_α−C(=O) bond in Asn16 were cut. Ala81 is modeled as the full residue forming peptide bonds to both Ala82 and Pro80, in which the C_α−C(=O) bonds were cut. Asn79 was cut between C_α and C_β. His85 was modeled as an imidazole ring cutting the C_α−C_β bond. Glu104 was cut between C_β and C_γ. Ser120 and Ser123 were modeled as water molecules, cutting the C_β−O(H) bonds. Moreover, a varying number of water molecules were included in the models. First, the crystal water molecule that stabilizes the enolate intermediate and that forms hydrogen bonds to Pro15, Asn16, and Ala82 in a number of crystal structures was included in all models. In M6, M7, and M8, a second water molecule was added by doubly protonating an oxygen atom of a buffer molecule (tartaric acid) that was resolved in the proximity of the active site. Model M7 is equivalent to M6 except that Glu104 is protonated in the former. In model M8, Glu104 and the two hydrogen-bonding water molecules modeling Ser120 and Ser123 are excluded. In the MD simulations, Tyr28 forms hydrogen bonds to Arg113 and to a water molecule. Furthermore, the carbonyl oxygen on C3 in DHQ is, in addition to the conserved water, hydrogen bonding to at least one more water molecule. Accordingly, we added two more water molecules in models M6W and M8W, which otherwise are identical to M6 and M8. One is hydrogen bonding to Tyr28 and is positioned in between the water molecule present in model M6 and Tyr28 (Figure 2). The second water molecule is hydrogen bonding to the carbonyl oxygen on C3 in DHQ and to the carbonyl in Pro15.

References

(1) Herrmann, K. M. The shikimate pathway as an entry to aromatic secondary metabolism. *Plant Physiol.* **1995**, *107*, 7.

(2) Herrmann, K. M. The shikimate pathway: Early steps in the biosynthesis of aromatic compounds. *Plant Cell* **1995**, *7*, 907.

(3) Herrmann, K. M.; Weaver, L. M. The shikimate pathway. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **1999**, *50*, 473.

(4) Schmid, J.; Amrhein, N. Molecular organization of the shikimate pathway in higher plants. *Phytochemistry* **1995**, *39*, 737.

(5) Steinrücken, H. C.; Amrhein, N. The herbicide glyphosate is a potent inhibitor of 5-enolpyruvyl-shikimic acid-3-phosphate synthase. *Biochem. Biophys. Res. Commun.* **1980**, *94*, 1207.

(6) Gourley, D. G.; Shrive, A. K.; Polikarpov, I.; Krell, T.; Coggins, J. R.; Hawkins, A. R.; Isaacs, N. W.; Sawyer, L. The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction. *Nat. Struct. Biol.* **1999**, *6*, 521.

(7) Shneier, A.; Harris, J.; Kleanthous, C.; Coggins, J. R.; Hawkins, A. R.; Abell, C. Evidence for opposite stereochemical courses for the reaction catalysed by type I and type II dehydroquinases. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 1399.

(8) Chaudhuri, S.; Duncan, K.; Graham, L. D.; Coggins, J. R. Identification of the active-site lysine residues of two biosynthetic 3-dehydroquinases. *Biochem. J.* **1991**, *275*, 1.

(9) White, P. J.; Young, J.; Hunter, I. S.; Nimmo, H. G.; Coggins, J. R. The purification and characterization of 3-dehydroquinase from *Streptomyces coelicolor*. *Biochem. J.* **1990**, *265*, 735.

(10) Moore, J. D.; Lamb, H. K.; Garbe, T.; Servos, S.; Dougan, G.; Charles, I. G.; Hawkins, A. R. Inducible overproduction of the *Aspergillus nidulans* pentafunctional AROM protein and the type-I and -II 3-dehydroquinases from *Salmonella typhi* and *Mycobacterium tuberculosis*. *Biochem. J.* **1992**, *287*, 173.

(11) Robinson, D. A.; Stewart, K. A.; Price, N. C.; Chalk, P. A.; Coggins, J. R.; Lapthorn, A. J. Crystal structures of *Helicobacter pylori* type II dehydroquinase inhibitor complexes: New directions for inhibitor design. *J. Med. Chem.* **2006**, *49*, 1282.

(12) Kleanthous, C.; Deka, R.; Davis, K.; Kelly, S. M.; Cooper, A.; Harding, S. E.; Price, N. C.; Hawkins, A. R.; Coggins, J. R. A comparison of the enzymological and biophysical properties of two distinct classes of dehydroquinase enzymes. *Biochem. J.* **1992**, *282*, 687.

(13) Harris, J. M.; Gonzalez-Bello, C.; Kleanthous, C.; Hawkins, A. R.; Coggins, J. R.; Abell, C. Evidence from kinetic isotope studies for an enolate intermediate in the mechanism of type II dehydroquinases. *Biochem. J.* **1996**, *319*, 333.

(14) Price, N. C.; Boam, D. J.; Kelly, S. M.; Duncan, D.; Krell, T.; Gourley, D. G.; Coggins, J. R.; Virden, R.; Hawkins, A. R. The folding and assembly of the dodecameric type II dehydroquinases. *Biochem. J.* **1999**, *338*, 195.

(15) Maes, D.; Gonzalez-Ramirez, L. A.; Lopez-Jaramillo, J.; Yu, B.; De Bondt, H.; Zegers, I.; Afonina, E.; Garcia-Ruiz, J. M.; Gulnik, S. Structural study of the type II 3-dehydroquinate dehydratase from *Actinobacillus pleuropneumoniae*. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 463.

(16) Roszak, A. W.; Robinson, D. A.; Krell, T.; Hunter, I. S.; Fredrickson, M.; Abell, C.; Coggins, J. R.; Lapthorn, A. J. The structure and mechanism of the type II dehydroquinase from *Streptomyces coelicolor*. *Structure* **2002**, *10*, 493.

(17) Frederickson, M.; Parker, E. J.; Hawkins, A. R.; Coggins, J. R.; Abell, C. Selective inhibition of type II dehydroquinases. *J. Org. Chem.* **1999**, *64*, 2612.

(18) Frederickson, M.; Roszak, A. W.; Coggins, J. R.; Lapthorn, A. J.; Abell, C. (1*R*,4*S*,5*R*)-3-Fluoro-1,4,5-trihydroxy-2-cyclohexene-1-carboxylic acid: The fluoro analogue of the enolate intermediate in the reaction catalyzed by type II dehydroquinases. *Org. Biomol. Chem.* **2004**, *2*, 1592.

(19) Kerbarh, O.; Bulloch, E. M.; Payne, R. J.; Sahr, T.; Rébeillé, F.; Abell, C. Mechanistic and inhibition studies of chorismate-utilizing enzymes. *Biochem. Soc. Trans.* **2005**, *33*, 763.

(20) Payne, R. J.; Peyrot, F.; Kerbarh, O.; Abell, A. D.; Abell, C. Rational design, synthesis, and evaluation of nanomolar type II dehydroquinase inhibitors. *ChemMedChem* **2007**, *2*, 1015.

(21) Toscano, M. D.; Payne, R. J.; Chiba, A.; Kerbarh, O.; Abell, C. Nanomolar inhibition of type II dehydroquinase based on the enolate reaction mechanism. *ChemMedChem* **2007**, *2*, 101.

(22) 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.

(23) Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **1996**, *60*, 1271.

(24) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577.

(25) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.

(26) 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* **1988**, *37*, 785.

(27) 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 C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(28) Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032.

(29) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. Ab initio study of ionic solutions by a polarizable continuum dielectric model. *Chem. Phys. Lett.* **1998**, *286*, 253.

(30) Mennucci, B.; Tomasi, J. Continuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries. *J. Chem. Phys.* **1997**, *106*, 5151.

(31) Blomberg, M. R. A.; Siegbahn, P. E. M.; Babcock, G. T. Modeling electron transfer in biochemistry: A quantum chemical study of charge separation in *Rhodobacter sphaeroides* and photosystem II. *J. Am. Chem. Soc.* **1998**, *120*, 8812.

(32) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* **2000**, *112*, 7374.

(33) Tewari, Y. B.; Goldberg, R. N.; Hawkins, A. R.; Lamb, H. K. A thermodynamic study of the reactions: {2-dehydro-3-deoxy-arabino-heptanoate 7-phosphate(aq)=3-dehydroquinate(aq) + phosphate(aq)} and {3-dehydroquinate(aq)=3-dehydroshikimate(aq) + H_2O(l)}. *J. Chem. Thermodyn.* **2002**, *34*, 1671.

(34) Hammond, G. S. A correlation of reaction rates. *J. Am. Chem. Soc.* **1955**, *77*, 334.

(35) Leffler, J. E. Parameters for the description of transition states. *Science* **1953**, *117*, 340.

(36) Krell, T.; Horsburgh, M. J.; Cooper, A.; Kelly, S. M.; Coggins, J. R. Localization of the active site of type II dehydroquinases. Identification of a common arginine-containing motif in the two classes of dehydroquinases. *J. Biol. Chem.* **1996**, *271*, 24492.

(37) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. AMBER9 University of California, San Francisco, 2006.

(38) 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**, *1781*, 26.

CT800480D

JCTC Journal of Chemical Theory and Computation

CBS-QB3 + VTST Study of Methyl *N*-Methylcarbamate + OH Gas-Phase Reaction: Mechanism, Kinetics, and Branching Ratios

Claudia Zavala-Oseguera[†] and Annia Galano*[,‡]

Departamento de Química, Universidad de Guanajuato, Noria Alta s/n C.P. 36050, Guanajuato, Gto. México, and Departamento de Química, Universidad Autónoma Metropolitana-Iztapalapa, San Rafael Atlixco 186, Col. Vicentina, Iztapalapa, C.P. 09340, México, D. F. México

Received February 4, 2009

Abstract: Different paths of reaction have been modeled, at high level of theory, accounting for the possible atmospheric fate of methyl *N*-methylcarbamate (MMC). The OH hydrogen abstractions from the methyl groups are predicted to account for almost 100% of the MMC + OH gas-phase reaction. The H abstraction from the methyl group at the N side of MMC was found to be the main path of reaction, with contributions to the overall reaction from 96% at 260 K to 89.2% at 400 K. Hydrogen abstractions from the other methyl group were identified as a secondary, but significant path. The Arrhenius activation energy, in the temperature range 260−400 K, was found to be close to zero, but slightly negative. The proposed value is (−0.10 ± 0.02) kcal/mol. The excellent agreement with the scarce experimental data available supports the reliability of the data reported here for the first time. Different IVTST-M-*H*/*G* schemes were tested for dynamic calculations. According to our results, including the Hessians and gradients, as few as four nonstationary points along the MEP is sufficient to achieve an excellent accuracy, provided that the data for these few points were obtained from high-quality levels of theory.

Introduction

Carbamates are versatile organic compounds with a wide range of applications. They are used as agrochemical agents with pesticide, fungicide, and herbicide activities.[1−4] They are also used as drug intermediates in the pharmaceuticals industry,[1,2,5] and in the polymer industry in the synthesis of polyurethane[1,2] and also in peptide syntheses.[5] In addition, among the various amine-protecting groups, carbamates are commonly used due to their chemical stability toward acids, bases, and hydrogenation.[6] Their general formula is $R_1N(H)C(O)OR_2$, where R_1 and R_2 can be alkyl or aryl groups.

Numerous studies have been devoted to the synthesis of carbamates; however, there is very scarce information of their fate after they are released into the environment. Carbamates with low molecular weights are volatile organic compounds (VOCs), and as for any other VOC their reaction with OH radicals is expected to be an important, or even the dominant, atmospheric loss process.[7,8]

Methyl *N*-methylcarbamate (MMC) is the simplest of this broad class of compounds, and that is why it has been selected for the present study. In an early work by Daly and Ziolkowski,[9] MMC was found to unimolecularly decompose in the gas phase, yielding methyl isocyanate and methanol in the temperature range 370−422 K. The mechanistic pathway of this reaction was explained in terms of a four-membered cyclic transition state. However, no further studies have been conducted in this direction.

In a recent work, Kwok et al.[10] reported the rate constants of OH gas-phase reactions with four alkyl carbamates at (296 ± 2) K. For MMC, the reported value is $(4.3 \pm 1.2) \times 10^{-12}$ cm^3 $molecule^{-1}$ s^{-1}. They proposed,[10] by analogy with thiocarbamates,[11] that the OH radical reactions with carbamates are expected to proceed by H-atom abstraction from

* Corresponding author e-mail: agalano@prodigy.net.mx.
[†] Universidad de Guanajuato.
[‡] Universidad Autónoma Metropolitana-Iztapalapa.

10.1021/ct9000679 CCC: $40.75
Published on Web 04/17/2009

the C−H bonds of the alkyl substituent groups and by initial addition to the N atom of the −N(H)C(O)O− moiety. They also found that the measured rate constants for the carbamates increase by replacing a methyl group by an ethyl group on both the ester site and the N atom. However, this increase is larger for the latter (2.4 vs 1.9 times larger). From these results, it can be inferred that if similar alkyl groups are bonded to the N and the O atoms, the contribution of H abstraction from the alkyl group on the N atom should be larger than that of the alkyl group on the ester group.

However, to our best knowledge, there is no available information on the branching ratios of these reactions. Actually, there is scarce information on branching ratios for VOCs reactions in general. Therefore, further efforts in this direction are needed since it has been established that branching ratios between different products in multichanneled reactions are as important as the overall rate of reaction, in terms of practical applications and in the understanding of the fundamental mechanisms of chemical reactions.[12−15]

In view of the above exposed arguments, the principal aim of this work is to model the different reactive pathways of the OH reactions with MMC and to estimate the proportion of their contribution to the overall reactions. For that purpose, the mechanism and the kinetics of this reaction have been studied as accurately as possible, using CBS-QB3 methodology for energies and variational transition state theory (VTST) for rate constants. In addition, a study of the temperature dependence of the rate coefficients was performed over the temperature range 260−400 K, and kinetic parameters are proposed for the first time.

Computational Details

All of the electronic structure calculations have been performed with the Gaussian 03 package of programs.[16] The high-level composite method CBS-QB3[17] has been used for all of the calculations. The complete basis set (CBS) models are compound methods that extrapolate to the CBS limit by using a $N - 1$ asymptotic convergence of MP2 pair energies calculated from pair natural orbital expansions.[17,18] They were developed to overcome the truncation of the basis sets, which is a major source of error in quantum mechanical calculations. CBS-QB3 is a five-step method that starts with B3LYP geometry and frequency calculations, followed by CCSD(T), MP4SDQ, and MP2 single-point calculations and a CBS extrapolation.[17] It also corrects for spin contamination.

Local minima and transition states were identified by the number of imaginary frequencies (NIMAG = 0 or 1, respectively). Intrinsic reaction coordinate (IRC)[19] calculations were carried out at B3LYP/CBSB7 level of theory to confirm that the transition states structures connect the proper reactants and products. The paths have been computed by following the Gonzalez−Schlegel steepest descent path,[20] in mass-weighted internal coordinates. Twenty points were modeled on each side of the saddle points, with a gradient step size of 0.05 amu$^{1/2}$ bohr. Thermodynamic corrections to the energy at 298.15 K were included in the determination of the energy barriers and of the heats of reaction, which have been reported in terms of Gibbs free energies.

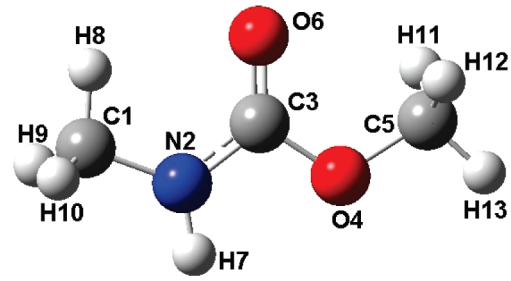


Figure 1. Atom numbering for methyl *N*-methylcarbamate.

The rate constant calculations for H abstraction paths were carried out by direct dynamics with the Interpolated Variational Transition-State Theory by Mapping (IVTST-M),[21] as implemented in POLYRATE 9.1.[22] In the IVTST-M algorithm, the calculations of rate constants evaluated by Canonical Variational Theory (CVT)[23] with Small-Curvature Tunneling (SCT) corrections are based on reaction-path data. The energies, energy gradients, and Hessians are computed at a small number of points along the minimum energy path (MEP) and fitted to splines under tension as functions of a mapped independent variable that is a nonlinear function of the reaction coordinate.[21] The notation IVTST-M-*H*/*G* means that interpolations are based on optimized calculations of stationary points (reactants, reactant complex, transition state, product complex, and product) plus *G* additional energies and gradients, and *H* additional Hessians corresponding to nonstationary points on the MEP. The different levels for dynamic calculations tested in this work are IVTST-M-8/10, IVTST-M-8/8, IVTST-M-6/8, IVTST-M-6/6, IVTST-M-4/10, IVTST-M-4/8, IVTST-M-4/6, IVTST-M-4/4, IVTST-M-2/4, and IVTST-M-2/2. All of the dynamic calculations in the present Article employ redundant internal coordinates for generalized normal-mode analyses.[25] We have used a step size of $0.01a_0$ for all IVTST-M calculations. The additional points on the MEP were chosen at a distance of $s = \pm 0.1$, 0.2, 0.4, and 0.6 amu$^{1/2}$ a_0 from the saddle point, where s represents the intrinsic reaction coordinate.

For OH addition and unimolecular fragmentation, the Conventional Transition State Theory (TST)[26] was used because the corresponding reaction barriers are high enough to safely assume that the variational effects are negligible. Tunneling factors were computed, in these cases, assuming an unsymmetrical, one-dimensional, Eckart function barrier.[27]

Results and Discussion

Structures and Energies. The optimized structure of MMC is shown in Figure 1, together with the atom numbering used through this Article. The atoms in the hydroxyl radical have been labeled as 14 and 15, referring to O and H atoms, respectively.

Seven different decomposition pathways related to MMC fate after it is released into the environment have been modeled.

Path I represents the unimolecular fragmentation proposed by Daly and Ziolkowski,[9] leading to the formation of methyl isocyanate and methanol:

$$CH_3N(H)C(O)OCH_3 \rightarrow CH_3NCO + CH_3OH \qquad (I)$$

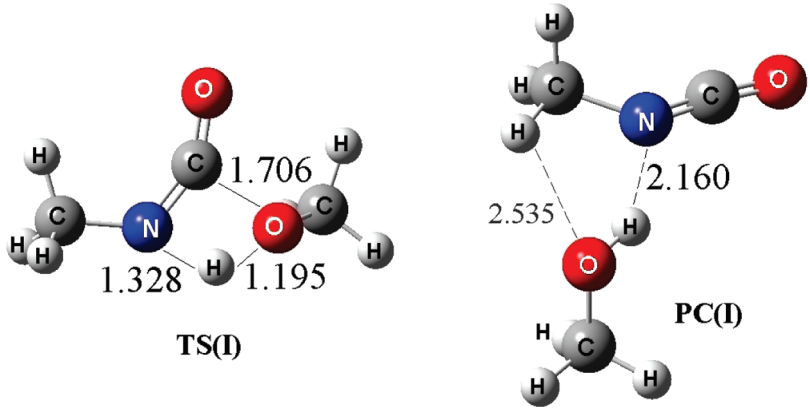


Figure 2. Optimized geometries of the transition state and product complex from path I.

A four-membered transition state (TS), consistent with the mechanism proposed by the above-mentioned authors, has been located for this path (Figure 2). It corresponds to a concerted mechanism with the breaking of C3−O4 and N2−H7 bonds and the formation of the O4−H7 bond occurring simultaneously. The bond distances d_{C3O4} and d_{N2H7} were found to be 0.344 and 0.321 Å larger in the TS than in MMC. The bond angles (a) in the four-membered ring have values of $a_{N2C3O4} = 90.4°$, $a_{C3O4H7} = 70.4°$, $a_{O4H7N2} = 116.8°$, and $a_{H7N2C3} = 82.2°$. While in MMC all non-hydrogen atoms are located on a plane, the dihedral/torsion angle is $t_{C5O4C3N2} = 118.9°$ in TS(I). A weak bonded complex has been identified in the product channel (Figure 2) with the main interaction involving the hydroxylic H in methanol and the N atom in methyl isocyanate. A secondary interaction was also found, between the O atom in methanol and one of the H atoms in methyl isocyanate.

The above-described transition state is 53.0 and 52.9 kcal/mol higher than the reactant, in terms of Gibbs free energy and enthalpy, respectively. This path was found to be endothermic and endergonic with $\Delta H_I = 24.5$, and $\Delta G_I = 12.5$ kcal/mol, at 298.15 K. According to these results, path I is not expected to significantly contribute to the overall environmental evolution of MMC.

Paths II, III, and IV account for OH hydrogen abstraction reactions from MMC. Paths II and IV represent H-atom abstraction from the methyl groups, as proposed by Kwok et al.,[10] while path III represents H abstraction from the NH group:

$$CH_3N(H)C(O)OCH_3 + {}^{\bullet}OH \rightarrow CH_3N(H)C(O)O{}^{\bullet}CH_2 + H_2O \quad \text{(II)}$$

$$\rightarrow CH_3N{}^{\bullet}C(O)OCH_3 + H_2O \quad \text{(III)}$$

$$\rightarrow {}^{\bullet}CH_2N(H)C(O)OCH_3 + H_2O \quad \text{(IV)}$$

Reactant complexes (RC) and product complexes (PC) were found for all of the abstraction paths in the entrance and exit channels, respectively (Figure 3). All of them show ring-like structures, caused by two interactions. In the RCs, the strongest interaction always involves the H atom in the hydroxyl radical and one of the O atoms in MMC. The secondary interaction involves the O in the OH radical and an H atom in the methyl (paths II and IV) or in the amide (path III) groups. In the PCs, the strongest interaction involves one H atom in the water fragment and one of the O atoms in the MMC formed radicals, while the secondary interaction involves the other H in water and the radical center in MMC radicals.

In all of the transition states, intramolecular H bonds were found between the H atom in the hydroxyl radical and one of the O atoms in MMC. This stabilizing interaction was found to be the strongest for TS(II) with the interaction distance $d'_{O6H14} = 2.186$ Å, followed by TS(IV) and TS(III), in that order, with d'_{O6H14} and d'_{O4H14} equal to 2.247 and 2.466 Å, respectively. The major structural changes in transition states II and IV, as compared to MMC, are the elongation of the C−H bond by 0.119 and 0.088 Å, respectively, and the shortening of the C−O and C−N distances by 0.025 and 0.019 Å. For TS(III), the main structural difference with respect to MMC is the elongation of the N−H distance by 0.155 Å.

All of the abstraction paths were found to be exothermic and exergonic with $\Delta H_{II} = -19.23$, $\Delta H_{III} = -10.9$, $\Delta H_{IV} = -24.5$, $\Delta G_{II} = -20.4$, $\Delta G_{III} = -11.6$, and $\Delta G_{IV} = -25.7$ kcal/mol, at 298.15 K. Accordingly, path IV is the most thermodynamically favored, while path III is the least favored, among the H abstraction channels. The barriers of reactions, in terms of Gibbs free energy, were found to be $\Delta G^{\ddagger}_{II} = 8.0$, $\Delta G^{\ddagger}_{III} = 11.0$, and $\Delta G^{\ddagger}_{IV} = 7.4$ kcal/mol. On the basis of these results, it is expected that H abstraction reactions mainly involve the methyl groups in MMC, and especially the one in the amidic end.

Three different pathways have been tested for OH addition reactions to MMC. Path V represents the OH addition to the N atom in the −N(H)C(O)O− moiety as proposed by Kwok et al.,[10] while paths VI and VII represent OH additions to the C and O atoms in the carbonyl group, respectively:

$$CH_3N(H)C(O)OCH_3 + {}^{\bullet}OH \rightarrow CH_3{}^{\bullet}N(H,OH)C(O)OCH_3 \quad \text{(V)}$$

$$\rightarrow CH_3N(H){}^{\bullet}C(O,OH)OCH_3 \quad \text{(VI)}$$

$$\rightarrow CH_3NHC{}^{\bullet}(O-OH)OCH_3 \quad \text{(VII)}$$

Any attempt to locate the products of additions V and VII invariably led to the separation into the reactant fragments. Product V evolves toward RC(III) structure, and product VII

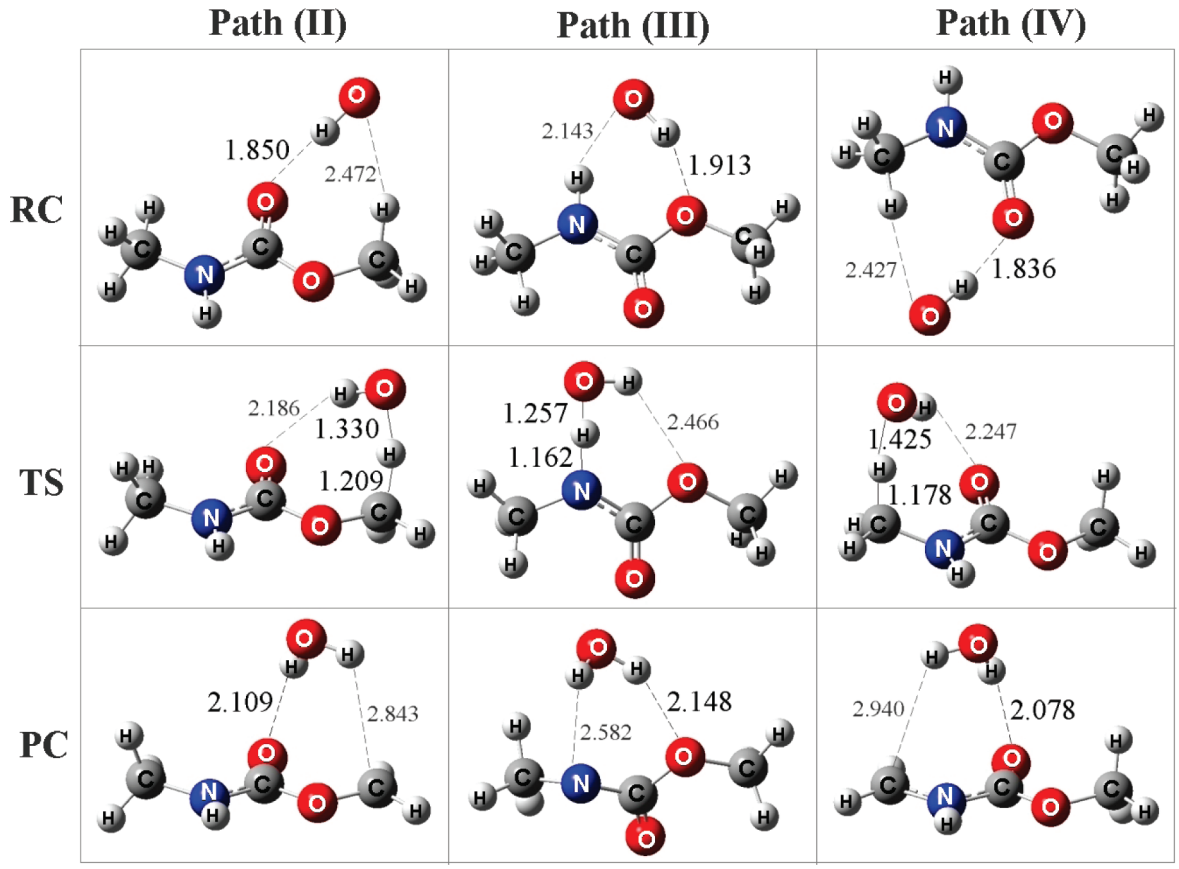


Figure 3. Optimized geometries of reactant complexes (RC), transition states (TS), and product complexes (PC) from paths II, III, and IV.

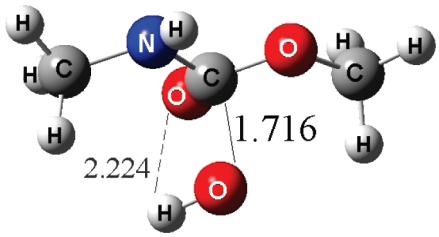


Figure 4. Optimized geometry of the transition state from path VI.

toward RC(II). Accordingly, the OH additions to the N atom of the −N(H)C(=O)O− moiety and to the O atom in the carbonyl group have been ruled out. For path VI, the addition product was located with C3 in tetrahedral configuration and d_{C3O14} = 1.408 Å, which corresponds to the newly formed C−O bond. A reactant complex was found in the entrance channel, and it is identical to RC(II). The structure shows a ring-like structure, with an intramolecular interaction between O6 in MMC and the H atom in the OH radical. The main structural changes, as compared to free MMC, are the elongation of the C3−O6 and C3−N2 bonds, by 0.085 and 0.023 Å, respectively.

Path VI (Figure 4) was found to be slightly exothermic and significantly endergonic, with $\Delta H_{VI} = -0.1$, $\Delta G_{VI} = 11.3$ kcal/mol. The calculated barrier height is 22.8 kcal/mol in terms of Gibbs free energy, at 298.15 K. According to these results, this path is not expected to significantly contribute to the overall OH + MMC reaction.

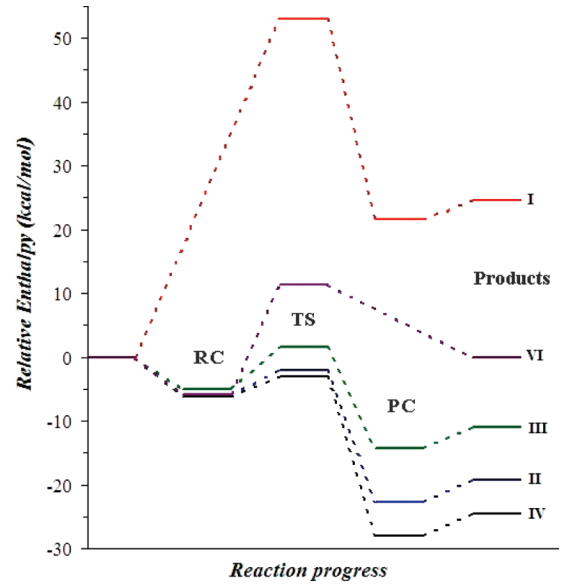


Figure 5. Energy profiles of the different studied paths of reaction.

The energy profiles of all of the computed paths of reactions in terms of enthalpy and relative to the isolated reactants are comparatively shown in Figure 5. The Gibbs free energy and enthalpy values are provided as Supporting Information (Table SI). As Figure 5 shows, the lower profiles correspond to H abstraction paths, which suggest that they are those contributing the most to the environmental loss process of MMC, when OH radicals are present.

Table 1. Calculated Rate Constants[a]

T (K)	k_I	k_{II}	k_{III}	k_{IV}	k_{VI}
260	6.33×10^{-29}	1.68×10^{-13}	8.03×10^{-16}	4.09×10^{-12}	4.08×10^{-25}
280	1.32×10^{-26}	1.66×10^{-13}	8.80×10^{-16}	4.05×10^{-12}	2.29×10^{-24}
298.15	1.20×10^{-24}	1.75×10^{-13}	9.63×10^{-16}	3.99×10^{-12}	8.97×10^{-24}
300	1.86×10^{-24}	1.79×10^{-13}	9.72×10^{-16}	3.98×10^{-12}	1.02×10^{-23}
320	1.84×10^{-22}	2.25×10^{-13}	1.08×10^{-15}	3.89×10^{-12}	3.81×10^{-23}
340	1.28×10^{-20}	2.73×10^{-13}	1.20×10^{-15}	3.80×10^{-12}	1.22×10^{-22}
360	6.34×10^{-19}	3.24×10^{-13}	1.35×10^{-15}	3.69×10^{-12}	3.46×10^{-22}
380	2.26×10^{-17}	3.74×10^{-13}	1.51×10^{-15}	3.59×10^{-12}	8.80×10^{-22}
400	5.98×10^{-16}	4.24×10^{-13}	1.69×10^{-15}	3.51×10^{-12}	2.05×10^{-21}

[a] I: TST, s^{-1}. II, III, and IV: IVTST-M-8/10, cm^3 $molecule^{-1}$ s^{-1}. VI: TST, cm^3 $molecule^{-1}$ s^{-1}.

Kinetics and Branching Ratios. As it is shown in Figure 5, paths I and VI have barriers that are higher than 10 kcal/mol. Accordingly, their kinetic study was performed only by conventional transition state theory (TST). For H abstraction paths, on the other hand, the IVTST-M-*H*/*G* was used.

For the calculation of the rate constants (k), a complex mechanism has been proposed for many radical−molecule reactions.[28−32,34] It involves a fast pre-equilibrium between the isolated reactants and a reactants complex (RC), followed by an irreversible step leading to the formation of products. In a classical treatment, the calculation of the overall rate coefficient depends only on the properties of reactants and transition states, and it can be shown that the RC energy and partition function cancels out in the rate constant expression.[35] However, when there is a possibility of quantum mechanical tunneling, the existence of the reactant complex might change the tunneling factor. In the present work, the H abstraction reactions have two reactants and two products, and weak bonded complexes have been found in both reactant and product channels. They have been explicitly considered in the potential surface and included in the IVTST-M calculations.

Because accurate rate constant calculations require proper evaluation of the partition functions (Q), the hindered rotor approximation has been used to correct the Q's corresponding to internal rotations with torsional barriers comparable to RT. Direct inspection of the low-frequency modes of the studied stationary points indicates that there are several modes that correspond to hindered rotations. These modes should be treated as hindered rotors instead as vibrations.[36] To make this correction, these modes were removed from the vibrational partition function of the corresponding specie and replaced with the hindered rotor partition function (Q^{HR}). In our calculations, we have adopted the analytical approximation to Q^{HR} for a one-dimensional hindered internal rotation proposed by Ayala and Schlegel.[37]

The values of the calculated rate constants are reported in Table 1. Those corresponding to paths II, III, and IV were obtained using the IVTST-M-8/10 methodology and SCT tunneling. Path I was found to be the slowest one, in the temperature range 260−400 K, despite the fact that its rate constant was calculated using TST, which represents an upper limit of CVT. Its rate constant shows a marked increase with temperature. Therefore, even though at temperatures relevant to atmospheric chemistry this path can be neglected, for higher temperatures it might become significant. The Arrhenius expression for this range was found to be $6.25 \times 10^8 \exp(-4421/RT)$. However, because this decomposition pathway was experimentally studied in the 370−422 K range, our data in an equivalent range were also fitted to an Arrhenius plot, and the activation energy was found to be 48.77 kcal/mol, in excellent agreement with the experimental value (48.06 kcal/mol).[9] This agreement supports the reliability of the calculated values.

Table 2. Branching Ratios (Γ), in Percentage, Corresponding to All of the Viable Paths (IVTST-M-8/10)

T (K)	Γ_I	Γ_{II}	Γ_{III}	Γ_{IV}	Γ_{VI}
260	1.7×10^{-21}	3.96	0.02	96.03	9.6×10^{-12}
280	3.6×10^{-19}	3.93	0.02	96.05	5.4×10^{-11}
298.15	3.3×10^{-17}	4.19	0.02	95.79	2.2×10^{-10}
300	5.1×10^{-17}	4.29	0.02	95.68	2.5×10^{-10}
320	5.1×10^{-15}	5.47	0.03	94.51	9.3×10^{-10}
340	3.6×10^{-13}	6.71	0.03	93.26	3.0×10^{-09}
360	1.8×10^{-11}	8.06	0.03	91.91	8.6×10^{-09}
380	6.6×10^{-10}	9.42	0.04	90.54	2.2×10^{-08}
400	1.7×10^{-08}	10.77	0.04	89.19	5.2×10^{-08}

The addition to the C atom in the carbonyl group was also found to be several orders of magnitude slower than those corresponding to H abstraction paths. Accordingly, it seems that the MMC reaction with OH radicals occurs mainly by H transfers from the carbamate to the radical. The largest rate constant was found for path IV in the whole studied temperature range, followed by that of path II. The H abstraction from the NH group was found to be the slowest among the H abstraction paths.

To quantify the contributions of the different paths to the overall rate coefficient, the corresponding branching ratios (Γ) expressed as percent were calculated and are reported in Table 2. They have been computed for a 24 h daytime average concentration of OH equal to 8.7×10^5 molecule cm^{-3}.[38] According to the values in this table, the only channels significantly contributing to the MMC + OH gas-phase reaction are the H abstractions from the methyl groups (paths II and IV). The contributions of all other paths together are less than 0.3% and can be neglected, at least within the studied temperature range. The contribution of H abstraction paths from amino groups to the overall rate coefficient in their reactions with OH radicals has been found to be up to 20%.[39] However, in this case, the contribution of path III was found to be negligible. This can be rationalized in terms of the structural differences between amine and −N(H)-C(O)O− moieties. The N atom in amines is bonded to alkyl groups, which have electron-donor character, helping to cope with the charge deficiency in the forming N radical center.

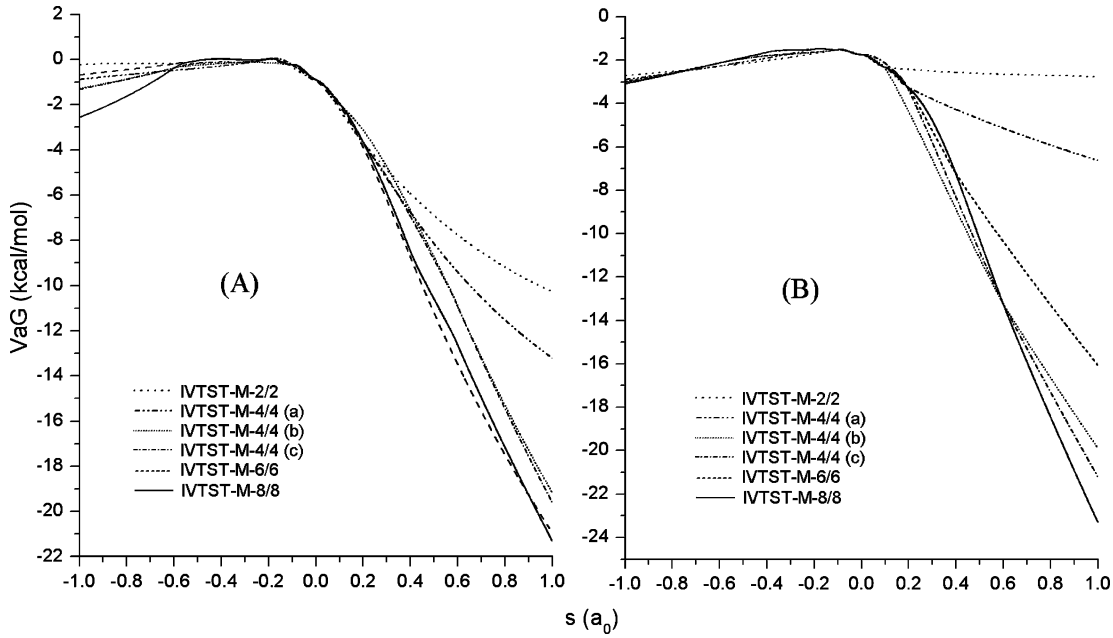


Figure 6. VaG(s), relative to isolated reactants, using the MEP generated by IVTST-M-2/2, IVTST-M-4/4 [H and G at $s = \pm 0.1$ and $\pm 0.2a_0$ (a); ± 0.1 and $\pm 0.6a_0$ (b); ± 0.2 and $\pm 0.6a_0$ (c)], IVTST-M-6/6, and IVTSTM-8/8. (A) Path II, (B) path IV.

The N atoms in carbamates, on the other hand, are bonded to −C(O)O− fragments with high electron-withdrawing character. Therefore, its presence deactivates the H abstraction from the NH site.

Path IV was found to contribute the most to the overall rate coefficient, from 96% at 260 K to 89.2% at 400 K, while the contribution of path II goes from ∼4% to 10.8%. These findings are in agreement with the information inferred from the work by Kwok et al.[10] that the contribution of H abstraction from the alkyl group on the N side should be larger than that of the alkyl group on the ester side, provided that similar alkyl groups are bonded to the N and the O atoms in carbamates. The presence of the electron-withdrawing −C(O)O− fragment in the vicinity of the methyl group, involved in the H abstraction, reduces the feasibility of such process, as compared to that of the equivalent alkyl group at the N side. In addition, it seems worthwhile to call attention to the fact that moving upward in the troposphere implies a lowering in temperatures; therefore, path II is expected to become less and less important as the altitude increases.

After the main paths of reaction were identified, the performance of different levels of dynamic calculations IVTST-M-*H*/*G* has been tested for them. Figure 6 shows the obtained ground-state vibrationally adiabatic potential energy paths defined as:

$$V_a^G(s) = V_{MEP}(s) + E_{int}(s) \quad (1)$$

where $V_{MEP}(s)$ is the classical potential energy path, and $E_{int}(s)$ is the local zero-point energy (ZPE) at s. Three different curves are shown within the IVTST-M-4/4 scheme. They differ in the s values at which the four nonstationary points have been chosen: (a) in the vicinity of the transition state, $s = \pm 0.1$ and $\pm 0.2a_0$; (b) two points in the vicinity ($s = \pm 0.1a_0$) and two points away ($s = \pm 0.6a_0$) from the TS; and (c) at the points where the IRC shows the maximum curvature. For path IV, the curvature on the product side shows one very sharp change at $s = +0.6a_0$, while the reactant side shows modest changes at $s = -0.2$ and $-0.6a_0$. For path II, there is also a sharp curvature change on the product side, this time at $s = +0.4a_0$, while on the reactant side the curvature change is almost negligible. As Figure 6 shows, the largest discrepancies among the different IVTST-M-*H*/*G* schemes appear on the product side and are more significant at large distances from the saddle points. For the main path of reaction IV, there is an excellent coincidence among the surfaces on the reactant side. The selection of the nonstationary points plays an important role in the shape of the interpolated VaG curve. If they are chosen taking into account the curvature of the MEP, the adiabatic surface obtained at IVTST-M-4/4 is very close to the IVTST-M-8/8 one. This is an interesting result because, depending on the size of the studied system, such a reduction in the number of computed Hessians might represent a significant saving in the computational cost. In the present work, the main paths of reaction show no tunneling effects; therefore, only slight differences are expected among the calculated rate constants as a consequence of the *H*/*G* choice. However, because the shape of the VaG curve plays a critical role in tunneling calculations, the above-discussed results show that a careful selection of the nonstationary points used for the interpolation is essential when tunneling becomes significant and a modest IVTST-M-*H*/*G* scheme is used. From this point on, and based on the results shown in Figure 6, the results reported with IVTST-M-4/4 refer to those obtained including the nonstationary points at the s values of maximum curvature. The absence of tunneling effects is caused by the low and broad barriers found in these cases.

Computing the overall rate coefficient as the sum of k_{II} and k_{IV}, which are the only paths significantly contributing to it, the nine tested IVTST-M schemes are very close and within the experimental error (Table 3), with the exception

Table 3. Overall Rate Coefficients at 298 K (k^{298}, in cm^3 molecule^{-1} s^{-1}), Arrhenius Activation Energies (*Ea*, in kcal/mol), Arrhenius Pre-exponential Factor (*A*, in cm^3 molecule^{-1} s^{-1}), and Reaction Coordinate of the Variational Transition States for Paths II and IV (*s*, in bohr)

	$k^{298} \times 10^{12}$	*Ea*	$A \times 10^{12}$	$s_{II}^{IVTST-M}$	$s_{IV}^{IVTST-M}$
IVTST-M-2/2	4.53	−0.16	3.46	−0.1872	−0.0989
IVTST-M-2/4	4.23	−0.06	3.83	−0.0832	−0.0836
IVTST-M-4/4	4.44	−0.15	3.47	−0.1758	−0.0900
IVTST-M-4/6	4.17	−0.10	3.50	−0.1921	−0.1837
IVTST-M-4/8	4.17	−0.11	3.45	−0.1919	−0.1831
IVTST-M-4/10	4.18	−0.11	3.48	−0.1955	−0.1835
IVTST-M-6/6	4.18	−0.09	3.57	−0.1888	−0.1798
IVTST-M-6/8	4.29	−0.12	3.47	−0.2030	−0.1904
IVTST-M-8/10	4.17	−0.12	3.40	−0.4855	−0.1812
exp[10]	4.30 ± 1.2				

of that obtained within the IVTST-M 2/2 scheme. However, even in this case, the calculated rate constant is only 1.05 times larger than the experimental one, which represents about 5.5% error. The Arrhenius activation energies were found to be slightly negative, varying from −0.06 to −0.16 kcal/mol. On the basis of the results from the largest values of *H* and *G*, a value of (−0.10 ± 0.02) kcal/mol is proposed for the activation energy. All of the calculated pre-exponential factors show very similar values, in the range (3.40−3.83) × 10^{-12} cm^3 molecule^{-1} s^{-1}. The variational transition state was found to be located on the reactant side for all of the tested IVTST-M-*H*/*G* schemes.

The values of the overall rate coefficients obtained using the IVTST-M-8/10 at the different studied temperatures together with the ratios of the values obtained within the other tested schemes with respect to IVTST-M-8/10 are reported in Table SII, as Supporting Information. The values of these ratios are all very close to unity, confirming the consistency of the IVTST-M-*H*/*G* methodology. As expected, the largest ratios correspond to the lowest *H* values (*H* = 2). The variational effects defined as k^{TST}/k^{ICVT} allow quantifying the overestimation of the TST rate constants, that is, the importance of using VTST. They were found to be rather small for path II: smaller than 1.8 regardless of the temperature and the *H*/*G* combination (Table SIII). For path IV, the variational effects are significantly larger, with values up to 7.02. Because both paths represent OH hydrogen abstractions from methyl groups, it seems that the atom directly bonded to the methyl group is responsible for the difference in the variational effects. Apparently, the presence of the O atom in the vicinity of the reaction site influences the thermal correction to Gibbs free energy to a larger extension than does the presence of an N atom, probably due to the relative higher electronegativity of the oxygen. It might lead to a tighter transition state with a larger loss of entropy, relative to the reactants. The variation of the k^{TST}/k^{ICVT} ratio was found to be almost independent of the IVTST-M-*H*/*G* level, for all of the tested *H*/*G* combinations, with the exception of IVTST-M-2/4 that led to variational effects for path II that are about one-half the magnitude of those obtained for any other tested *H*/*G* combination. In all cases, and for both paths, the variational effects were found to decrease as the temperature rises, which is consisted with the results reported for the OH hydrogen abstraction from methane.[21]

In the original paper presenting IVTST-M-*H*/*G*, it was established that using $H \geq 6$ is enough to obtain good results.[21] Considering the data from Tables 3, SII, and SIII all together, it seems that providing Hessians and gradients for as few as four nonstationary points along the MEP is sufficient to achieve an excellent accuracy, provided that the data for these few points are obtained at high-quality levels of theory and that they are chosen taking the curvature of the surface into account. In fact, it seems that the interpolation by mapping works efficiently even when only four Hessians and four gradients are explicitly provided, at least for H abstraction reactions from alkyl groups.

Conclusions

Different paths of reaction have been modeled, at high level of theory, accounting for the possible atmospheric fate of methyl *N*-methylcarbamate. It has been shown that its unimolecular gas-phase decomposition yielding methyl isocyanate and methanol is negligible, especially at atmospheric temperatures and when OH radicals are present. The addition of the OH radical to the N atom in the −N(H)C(O)O− moiety has been ruled out as well. The OH hydrogen abstractions from the methyl groups in MMC account for almost the 100% of the MMC + OH gas-phase reaction. The H abstraction from the methyl group at the N side was found to be the main path of reaction, with contributions to the overall reaction from 96% at 260 K to 89.2% at 400 K. Hydrogen abstractions from the other methyl group were identified as a secondary, but significant path, with contributions from ∼4% to 10.8%. The Arrhenius activation energy, in the temperature range 260−400 K, was found to be close to zero, but slightly negative. The proposed value is (−0.10 ± 0.02) kcal/mol. The excellent agreement with the scarce experimental data available supports the reliability of the data reported here for the first time. Different IVTST-M-*H*/*G* schemes were tested for dynamic calculations. The high efficiency of the Interpolated Variational Transition-State Theory by Mapping, when only a very small number of points along the reaction path are explicitly included in the calculations, has been proven. According to our results, providing the Hessians and gradients for as few as four nonstationary points along the MEP is sufficient to achieve an excellent accuracy, provided that the data for these few points were obtained from high-quality levels of theory, at least for H abstraction reactions from alkyl groups. It should be noticed, however, that for reactions with significant tunneling the selection of the nonstationary points plays an important role in the quality of the interpolated MEP. For better results, those points at *s* values corresponding to the maximum curvature should be included.

Acknowledgment. We thank Prof. Donald G. Truhlar for providing the POLYRATE 9.1 program and Prof. J. R. Alvarez-Idaboy for helpful discussions. A.G. also thanks professors D. G. Truhlar and J. Zheng for their valuable help and advice on dynamic calculations, and Laboratorio de

Visualización y Cómputo Paralelo at UAM - Iztapalapa for the access to its computer facilities. The work in Guanajuato was funded by Conacyt (Grant 57892). C.Z.-O. acknowledges CONACyT for the Master fellowship.

Supporting Information Available: Enthalpies and Gibbs free energies of the stationary points, relative to the isolated reactants. Overall IVTST-M-8/10 rate coefficients at the different studied temperatures and ratios IVTST-M-*H*/*G*/IVTST-M-8/10. Variational effects k^{TST}/k^{IVTST} within the 260−400 K temperature range for paths II and IV, from all tested IVTST-M-*H*/*G* methods. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dibenedetto, A.; Aresta, M.; Fragale, C.; Narracci, M. *Green Chem.* **2002**, *4*, 439.

(2) Gupta, S. P.; Shivarkar, A. B.; Chaudhari, R. V. *J. Chem. Soc., Chem. Commun.* **2001**, 2620.

(3) Motolcsy, G.; Nadasy, M.; Andriska, V. *Pesticide Chemistry*; Academiai Kiado: Budapest, 1988; p 90.

(4) Thompson, A. *Pesticide Outlook* **2002**, *13*, 84.

(5) Lawrence, M.; Klein, D. L.; Nemotho, P. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 157.

(6) Greene, W. T.; Wuts, P. G. M. *Protective Groups in Organic Synthesis*, 2nd ed.; Wiley: New York, 1991; pp 327 and 403.

(7) Atkinson, R. Atmospheric Transformations of Automotive Emissions. In *Air Pollution, the Automobile, and Public Health*; Watson, A. Y., Bates, R. R., Kennedy, D., Eds.; National Academy Press: Washington, DC, 1988; pp 99−132.

(8) Atkinson, R. J. *Phys. Chem. Ref. Data* **1994**, 1, Monograph 2.

(9) Daly, N. J.; Ziolkowski, F. *Aust. J. Chem.* **1972**, *25*, 1453.

(10) Kwok, E. S. C.; Aschmann, S.; Atkinson, R. *Environ. Sci. Technol.* **1996**, *30*, 329.

(11) Kwok, E. S. C.; Atkinson, R.; Arey, J. *Environ. Sci. Technol.* **1992**, *26*, 1798.

(12) Seakins, P. W. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2007**, *103*, 173.

(13) Butkovskaya, N. I.; Kukui, A.; Le Bras, G. *J. Phys. Chem. A* **2004**, *108*, 1160.

(14) Butkovskaya, N. I.; Setser, D. W. *J. Phys. Chem. A* **1999**, *103*, 6921.

(15) Butkovskaya, N. I.; Kukui, A.; Pouvesle, N.; Le Bras, G. *J. Phys. Chem. A* **2004**, *108*, 7021.

(16) 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 E.01; Gaussian, Inc.: Wallingford, CT, 2004.

(17) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822.

(18) (a) Nyden, M. R.; Petersson, G. A. *J. Chem. Phys.* **1981**, *75*, 1843. (b) Al-Laham, M. A.; Petersson, G. A. *J. Chem. Phys.* **1991**, *94*, 6081. (c) Petersson, G. A.; Tensfeldt, T. G.; Montgomery, J. A. *J. Chem. Phys.* **1991**, *94*, 6091. (d) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery, J. A.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 10570. (e) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532.

(19) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.

(20) (a) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154. (b) Gonzalez, C.; Schlegel, H. B. *J. Phys. Chem.* **1990**, *94*, 5523.

(21) Corchado, J. C.; Coitiño, E. L.; Chuang, Y.-Y.; Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 2424.

(22) Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Villà, J.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Melissas, V. S.; Lynch, B. J.; Rossi, I.; Coitiño, E. L.; Fernandez-Ramos, A.; Pu, J.; Albu, T. V.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Truhlar, D. G. *POLYRATE-version 9.1*; University of Minnesota: Minneapolis, 2002.

(23) (a) Isaacson, A. D.; Truhlar, D. G. *J. Chem. Phys.* **1982**, *76*, 1380. (b) Truhlar, D. G.; Garrett, B. C. *Annu. Rev. Phys. Chem.* **1984**, *35*, 159. (c) Chuang, Y.-Y.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **1998**, *70*, 887.

(24) (a) Garrett, B. C.; Truhlar, D. G.; Grev, R. S.; Magnuson, A. W. *J. Phys. Chem.* **1980**, *84*, 1730. (b) Truhlar, D. G.; Garrett, B. C. *Annu. Rev. Phys. Chem.* **1984**, *35*, 159. (c) Truhlar, D. G.; Liu, Y.-P.; Schenter, G. K.; Garrett, B. C. *J. Phys. Chem.* **1994**, *98*, 8396.

(25) Chuang, Y.-Y.; Truhlar, D. G. *J. Phys. Chem. A* **1998**, *102*, 242.

(26) (a) Eyring, H. *J. Chem. Phys.* **1935**, *3*, 107. (b) Truhlar, D. G.; Hase, W. L.; Hynes, J. T. *J. Phys. Chem.* **1983**, *87*, 2664.

(27) Eckart, C. *Phys. Rev.* **1930**, *35*, 1303.

(28) Singleton, D. L.; Cvetanovic, R. J. *J. Am. Chem. Soc.* **1976**, *98*, 6812.

(29) Mora-Diez, N.; Alvarez-Idaboy, J. R.; Boyd R, J. *J. Phys. Chem. A* **2001**, *105*, 9034.

(30) Alvarez-Idaboy, J. R.; Mora-Diez, N.; Boyd R, J.; Vivier-Bunge, A. *J. Am. Chem. Soc.* **2001**, *123*, 2018.

(31) Galano, A.; Alvarez-Idaboy, J. R.; Ruiz-Santoyo, M. E.; Vivier-Bunge, A. *ChemPhysChem* **2004**, *5*, 1379.

(32) Francisco-Marquez, M.; Alvarez-Idaboy, J. R.; Galano, A.; Vivier-Bunge, A. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1392.

(33) Galano, A.; Alvarez-Idaboy, J. R.; Ruiz-Santoyo, M. E.; Vivier-Bunge, A. *J. Phys. Chem. A* **2005**, *109*, 169.

(34) Galano, A.; Alvarez-Idaboy, J. R. Atmospheric Reactions of Oxygenated Compounds + OH radicals: Role of Hydrogen-Bonded Intermediates and Transition States. In *Advances in Quantum Chemistry: Applications of Quantum Chemistry*

to the Atmosphere; Goodsite, M. E., Johnson, M. S., Eds.; Elsevier Pub.: Amsterdam, 2008; Chapter 12, pp 245−274.

(35) Alvarez-Idaboy, J. R.; Cruz-Torres, A.; Galano, A.; Ruiz-Santoyo, M. E. *J. Phys. Chem. A* **2004**, *108*, 2740.

(36) Jacox, M. E. *Vibrational and Electronic Energy Levels of Polyatomic Transient Molecules*; NIST: Gaithersburg, MD, 1998; Vol. 69, p 945.

(37) Ayala, P. Y.; Schlegel, H. B. *J. Chem. Phys.* **1998**, *108*, 2314.

(38) Prinn, R.; Cunnold, D.; Simmonds, P.; Alyea, F.; Boldi, R.; Crawford, A.; Fraser, P.; Gutzler, D.; Hartley, D.; Rosen, R.; Rasmussen, R. *J. Geophys. Res.* **1992**, *97*, 2445.

(39) Galano, A.; Alvarez-Idaboy, J. R. *J. Chem. Theory Comput.* **2008**, *4*, 322.

CT9000679

JCTC Journal of Chemical Theory and Computation

Comparison of Three Perturbation Molecular Dynamics Methods for Modeling Conformational Transitions

He Huang,[†] Elif Ozkirimli,[†,‡] and Carol Beth Post*[,†]

Department of Medicinal Chemistry and Molecular Pharmacology, Markey Center for Structural Biology and Purdue Cancer Center, Purdue University, West Lafayette, Indiana 47907, and Department of Chemical Engineering, Bogazici University, Bebek, 34342, Turkey

Received January 9, 2009

Abstract: Targeted, steered, and biased molecular dynamics (MD) are widely used methods for studying transition processes of biomolecules. They share the common feature of adding external perturbations along a conformational progress variable to guide the transition in a predefined direction in conformational space, yet differ in how these perturbations are applied. In the present paper, we report a comparison of these three methods on generating transition paths for two different processes: the unfolding of the B domain of protein A and a conformational transition of the catalytic domain of a Src kinase Lyn. Transition pathways were calculated with different simulation parameters including the choice of progress variable and the simulation length or biasing force constant. A comparison of the generated paths based on structural similarity finds that the three perturbation MD methods generate similar transition paths for a given progress variable in most cases. On the other hand, the path depends more strongly on the choice of progress variable used to move the system between the initial and final states. Potentials of mean force (PMF) were calculated starting from unfolding trajectories to estimate the relative probabilities of the paths. A lower PMF was found for the lowest biasing force constant with BMD.

1. Introduction

Functionally important dynamic processes of proteins, such as folding/unfolding and allosteric conformational transitions, occur on the microsecond to millisecond time scale. Molecular dynamics simulation is a useful tool to elucidate the atomistic detail of protein dynamics; however, simulations of all-atom models are limited to the submicrosecond time range. To overcome this time scale problem, methods that utilize the principles of molecular dynamics, with some external perturbations to accelerate the reaction and guide the system toward a target state, have been developed.[1−6] Here, such methods are collected under the title of "perturbation molecular dynamics". They aim to identify possible transition pathways as well as energy barriers and metastable intermediates. Such pathways can then be further examined by thermodynamic simulation methods.

The most commonly used perturbation molecular dynamics methods are targeted molecular dynamics (TMD), steered molecular dynamics (SMD), and biased molecular dynamics (BMD). These methods share the common feature of guiding the transition between two end states through some progress variable (reaction coordinate), though they differ in the way that the progress variable is controlled. As originally introduced by Schlitter,[1,2] the TMD methodology imposes a time-dependent holonomic constraint on the rmsd to a target structure. SMD simulations were first used by Grubmueller[3] and Leech,[4] and were widely applied shortly after by Schulten et al.[7−10] SMD is akin to atomic force microscopy (AFM) in that a harmonic restraint based on a reference point moves the system toward the target when the reference point is updated. We note that in some later publications[11−16] the term TMD was associated with the rmsd progress variable

* Corresponding author e-mail: cbp@purdue.edu.
[†] Purdue University.
[‡] Bogazici University.

10.1021/ct9000153 CCC: $40.75
Published on Web 04/09/2009

with harmonic restraint rather than holonomic constraint. In the present paper, we denote the holonomic constraint as TMD and the harmonic restraint as SMD to distinguish the perturbation form regardless of the progress variable. BMD, also known as the adiabatic bias molecular dynamics, was originally proposed independently by Marchi and Ballone[5] and Paci and Karplus.[6] It provides the least perturbation among the three in that the system feels no force if it moves toward the target and the biasing potential is nonzero only if the system moves away from the target.

These three methods are commonly used to study transition processes of biomolecules[11−30] They are also used to complement AFM experiments that examine the mechanical properties of macromolecules.[10,31] Given the interest in these perturbation MD methods, an assessment of the effect of the choice of perturbation method, progress variable, and other simulation parameters on the resulting transition paths is of value.

In the present paper, we report a systematic comparison of the conformations and energetics of trajectories generated by the three perturbation molecular dynamics methods for two transition processes. The first example system is the unfolding of the B domain of staphylococcal protein A (BdpA). BdpA is a three-helix-bundle protein for which folding/unfolding has been studied extensively by experiments[32−34] and computer simulations.[35−38] Its unfolding is an example where the progress variable does not define a single target configuration. The second process is the conformational change between active and inactive structures of the kinase catalytic domain (CD) of Lyn, a member of the Src family of protein tyrosine kinases. This transition is an example that has defined target configurations at both end states. For TMD, SMD, and BMD paths generated for both systems, we examined the effect of different simulation conditions, including the choice of progress variable and perturbation strength. Our results suggest that, for the most part, the three perturbation MD methods generate similar transition paths for a given progress variable even though the time dependence of the progress variables differs substantially. On the other hand, the path depends strongly on the choice of the progress variable.

2. Methods

2.1. Three Perturbation MD Methods. Targeted molecular dynamics (TMD) introduces the most constrained perturbation among the three methods. While the other two methods add restraining potentials to guide the system, TMD imposes a holonomic constraint onto the dynamics of the system:[2]

$$\phi(\rho(x(t)), \rho_0(t)) = 0 \tag{1}$$

where $\rho(x)$ is the progress variable defined as a function of the coordinate, x, $\rho_0(t)$ is the reference value of ρ at time t, and ϕ is a function of ρ and ρ_0 which equals zero when $\rho = \rho_0$; for example

$$\phi(x) = \rho - \rho_0 \tag{2}$$

The constraint adds onto the system a constraining force

$$F^C = \lambda \nabla_x \phi \tag{3}$$

which keeps the progress variable ρ following the reference value $\rho_0(t)$ exactly. Here λ is a Lagrange parameter determined according to eq 1, and the reference value $\rho_0(t)$ moves at a constant rate v toward the target value:

$$\rho_0(t) = \rho_0(t_0) + v(t - t_0) \tag{4}$$

Steered molecular dynamics (SMD) corresponds closely to micromanipulation by AFM[39] when it uses a single interatomic distance as the progress variable. Computationally, it adds a full harmonic potential to restrain the progress variable around a reference value, which is moved to the target value at a constant rate v:

$$H(\rho) = \frac{\alpha}{2}(\rho - \rho_0)^2 \tag{5}$$

$$\rho_0(t) = \rho_0(t_0) + v(t - t_0) \tag{6}$$

where H is the biasing potential, α is the force constant, ρ is the progress variable, and ρ_0 is the reference point.

Biased molecular dynamics (BMD) guides the change of the progress variable by penalizing a move in the undesired direction through a one-sided harmonic potential. At each time step, the reference point is updated to the previously sampled value that is closest to the target. The method is defined by the following equations assuming the system is moved in the direction in which the progress variable ρ increases:

$$H(\rho) = \begin{cases} \frac{\alpha}{2}(\rho - \rho_0)^2 & (\rho < \rho_0) \\ 0 & (\rho \geq \rho_0) \end{cases} \tag{7}$$

$$\rho_0(t) = \begin{cases} \rho_0(t - \Delta t) & (\rho < \rho_0) \\ \rho(t) & (\rho \geq \rho_0) \end{cases} \tag{8}$$

where Δt is the simulation time step. Among the three methods, BMD provides the least restrained perturbation to the molecular system in that progress in the direction toward the target occurs without external perturbation.

2.2. Molecular Systems. The molecular system of BdpA was derived from the NMR structure (PDB ID: 1BDD).[40] The C- and N-terminal loops were removed, and residues 10−55 were kept. A 2-ns equilibrium MD simulation was calculated with the CHARMM22 force field and the GBSW implicit solvent model[41] in CHARMM.[42] Four configurations from the equilibrium run separated by 500 ps were saved and used as initial structures of the perturbation MD runs. The CHARMM22 force field and GBSW solvation model were used in all perturbation MD runs for BdpA unfolding.

For the Lyn CD, the active and inactive coordinates were obtained by homology modeling based on crystallographic structures of the Lck kinase (PDB ID: 3LCK)[43] and Hck kinase (PDB ID: 1QCF),[44] respectively. Both structures were equilibrated with the CHARMM22 force field with approximately 9400 TIP3P waters in a periodic rhombic dodecahedral box at 298 K for 300−400 ps, as previously reported.[23] The CHARMM22 force field with explicit TIP3P water was also used in all perturbation MD runs for the Lyn CD conformational transition.

2.3. Progress Variables. For BdpA unfolding, the three perturbation MD methods were examined with two progress variables, either the end-to-end distance between the two

Table 1. Summary of BdpA Unfolding Trajectories

progress variable	method	length (ns)	force constant (kcal/(mol·Å²))	rate (Å/ps)	no.[a]	code
h		0.5	–	0.1	2	hT1
	TMD	2	–	0.025	2	hT2
		10	–	0.005	2	hT3
		0.5	50	0.1	2	hS1
	SMD	2	50	0.025	2	hS2
		10	50	0.005	2	hS3
		0.5	50	–	2	hB1
	BMD	2	20	–	2	hB2
		10	8	–	2	hB3
R_g		0.5	–	0.05	2	RT1
	TMD	2	–	0.0125	2	RT2
		10	–	0.0025	4	RT3
		0.5	5000	0.05	2	RS1
	SMD	2	5000	0.0125	2	RS2
		10	5000	0.0025	4	RS3
		0.5	5000	–	2	RB1
	BMD	2	400	–	2	RB2
		10	110	–	4	RB3

[a] Number of independent trajectories that were calculated.

terminal C_α atoms ($\rho = h$) or the radius of gyration based on all heavy atoms ($\rho = R_g$), which is defined as

$$R_g = \sqrt{\frac{1}{N}\sum_i^N \left| r_i - \frac{1}{N}\sum_j^N r_j \right|^2} \tag{9}$$

where r_i is the Cartesian coordinate of the ith heavy atom and N equals the total number of heavy atoms. Each combination of method and progress variable was used with three different perturbation strengths as reflected by the simulation lengths: 0.5, 2, and 10 ns. With TMD or SMD, the simulation length was controlled directly by the rate for updating the reference point, while in BMD, in which no such parameter is available, it was controlled by tuning the force constant so that the unfolding process happens in roughly the specified time. For each set of simulation conditions (perturbation method, progress variable, and simulation length), two to four (see Table 1 for details) trajectories were calculated with different initial coordinates and velocities.

The Lyn CD conformational transition for both activation (inactive to active CD) and deactivation (active to inactive) was simulated by all three perturbation MD methods with the progress variable mean square internal deviation (ρ = MSID) defined in terms of internal distances:

$$\mathrm{MSID} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j>i}^{N}(d_{ij} - d_{ij}^0)^2 \tag{10}$$

where d_{ij} and d_{ij}^0 are distances between heavy atoms i and j in the current and target structures, respectively, and N is the number of atoms. This progress variable was originally used to investigate partially unfolded protein in combination with BMD by Paci et al.,[45] and later applied to Src kinase activation.[23] A similar progress variable defined upon internal distances was used by Markwick et al. together with a mass weighting.[46]

A second progress variable, root-mean-square deviation (ρ = rmsd), was examined with only the TMD method, and is defined as

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|r_i - r_i^0|^2} \tag{11}$$

where r_i and r_i^0 are Cartesian coordinates of the ith heavy atom in the superimposed current and target structures, respectively, and N is the total number of heavy atoms. The combination of rmsd as the progress variable and the holonomic constraint as the perturbation method coincides with the original TMD method used by Schlitter et al.[1,2] For each combination of method, progress variable, and direction, two to four (see Table 2 for details) trajectories were calculated with different updating rates, v, of the progress variable (TMD, SMD) or force constants, α (BMD).

2.4. Trajectory Averaged rmsd. The structural similarity of the BdpA unfolding trajectories was evaluated from a trajectory averaged rmsd calculated to measure the pairwise similarity between different trajectories. For each unfolding trajectory with either h or R_g as the progress variable, snapshots were binned according to R_g into 0.5-Å-wide bins. An average configuration $\bar{X}_k$ was calculated for each bin k to represent the snapshots within that bin. Snapshots within each bin have similar structures, as indicated by the average rms fluctuation of 2.0, 2.0, and 1.6 Å for TMD, SMD, and BMD 10 ns trajectories, respectively. The within-bin variances for shorter trajectories are expected to be even lower. Between trajectories i and j, the trajectory averaged rmsd is defined as

$$\overline{\mathrm{rmsd}}_{ij} = \frac{1}{M}\sum_{k=1}^{M}\mathrm{rmsd}(\bar{X}_k^i, \bar{X}_k^j) \tag{12}$$

where M is the number of bins and rmsd is calculated after superimposing all heavy atoms. This single value provides a metric for the overall similarity between two trajectories. Because the rmsd between two extended structures is always small, we ignored snapshots with R_g greater than 17 Å in this calculation.

Table 2. Summary of Lyn CD Activation/Deactivation Trajectories

progress variable	method	direction[a]	length (ns)	force constant (kcal/(mol·Å⁴))	rate (Å/ps)	color[b]
MSID	TMD	A → I	0.16	–	0.1	black
			1.0	–	0.016	red
		I → A	0.16	–	0.1	green
			1.0	–	0.016	blue
	SMD	A → I	0.16	1000	0.1	black
			1.0	1000	0.016	red
		I → A	0.16	1000	0.1	green
			1.0	1000	0.016	blue
	BMD	A → I	1.9	1000	–	black
			1.5	2000	–	red
			1.5	3000	–	yellow
		I → A	1.3	5000	–	brown
			1.9	1000	–	green
			1.9	2000	–	blue
			1.7	3000	–	gray
			1.0	5000	–	purple
rmsd	TMD	A → I	0.2	–	0.03	black
			1.0	–	0.006	red
		I → A	0.2	–	0.03	green
			1.0	–	0.006	blue

[a] A, active state; I, inactive state. [b] Coloring for Figure 8a.

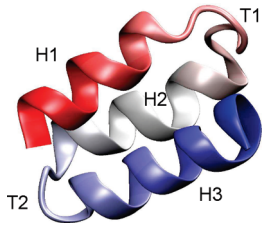


Figure 1. The native structure of BdpA[40] shown in cartoon representation. Helices H1, H2, and H3 are colored in red, white, and blue, respectively. This figure was generated with Visual Molecular Dynamics (VMD).[49]

2.5. Potential of Mean Force Calculation. The potential of mean force (PMF) was calculated with an initial path defined by a R_g-perturbed BdpA unfolding trajectory by using umbrella sampling[47] and the weighted histogram analysis method (WHAM).[48] For each trajectory, 41 snapshots were taken as the initial coordinates for umbrella sampling by choosing coordinates with R_g values closest to an equally spaced R_g series ranging from 9.75 to 15.75 Å in increments of 0.15 Å. Each of the 41 umbrella windows was simulated for 400 ps using a harmonic umbrella potential with force constant 10 kcal/(mol·Å^2) to restrain R_g around the initial value. The last 200 ps of the sampling was analyzed by WHAM to reconstruct the PMF profile with respect to R_g. The effect of initial-coordinate bias for a given trajectory was examined by choosing a different set of initial coordinates with a shifted R_g series ranging from 9.83 to 15.83 Å with 0.15 Å increments for the PMF calculations. Results for two trajectories showed that very similar PMF curves are obtained for each trajectory, indicating that the calculated PMF curve characterizes a trajectory rather than the specific selection of snapshots (data not shown).

3. Results

Tables 1 and 2 summarize the perturbation MD simulations carried out for the two transition systems. All simulations and analyses were carried out with the molecular dynamics program CHARMM.[42]

3.1. BdpA Unfolding. BdpA is a three-helix-bundle protein with a highly symmetrical topology. The three helices (H1, H2, and H3) of comparable length are joined by two turns (T1 and T2) in an antiparallel alignment to form two helix−turn−helix motifs (see Figure 1).

As listed in Table 1, a total of 42 unfolding trajectories of BdpA were generated with three perturbation MD methods (BMD/SMD/TMD), two progress variables (end-to-end distance h/radius of gyration R_g), and three different simulation lengths (0.5 ns/2 ns/10 ns). Figure 2 shows the time profiles of the progress variables for representative trajectories generated by the three methods. With TMD, the progress variables scale linearly with time as expected from the holonomic constraint. SMD also generated nearly linear progression curves due to the linear updating of the reference point, but the actual progress variables fluctuate about the linear line. With BMD, the dynamics of the progress variables appear more similar to natural fluctuation in the sense that their values change nonlinearly over time. For example, in the R_g-perturbed BMD simulation (Figure 2, right bottom panel), the progress variable R_g increases slowly for the first 8 ns until it reaches ∼11.5 Å, after which it rises with a much steeper slope. This nonlinear progression can be utilized to identify possible free energy barriers, because a barrier impedes spontaneous fluctuations along the progress variable and thus longer sampling time is needed to cross it.[20,21]

Effect of Perturbation Methods and Progress Variable on Determining Path. R_g and h were used to guide the unfolding transition of BdpA. Unfolding trajectories, from TMD, SMD and BMD, were compared by the trajectory averaged rmsd values ($\overline{\text{rmsd}}_{ij}$), for which the snapshots from each trajectory were binned according to their R_g values. The calculated unfolding trajectories were thus examined as paths connecting conformations in space rather than time evolution of the system. To compare two trajectories, the rmsd was calculated between average structures from each R_g bin, and the rmsd values were averaged over all bins (see Methods for details). The resulting $\overline{\text{rmsd}}_{ij}$ value measures the overall spatial similarity between two trajectories, and the all-against-all evaluation is plotted in the matrix in Figure 3 according to the code shown in Table 1. This pairwise similarity matrix shows that the 42 trajectories naturally fall into three clusters. Interestingly, all h-perturbed trajectories fall into cluster A, all R_g-perturbed trajectories except RB3 fall into cluster B, and all R_g-perturbed, RB3 trajectories fall into cluster C. The groups of trajectories are clearly distinct from each other: The $\overline{\text{rmsd}}_{ij}$ value within any cluster is around 4 Å while that between two different clusters averages near 11 Å.

It is apparent from the distinct clustering in Figure 3 that, for the case of BdpA unfolding, the global geometry of the transition path is largely determined by the progress variable being h or R_g, and not by the perturbation method. Comparison of trajectories within a cluster provides no evidence that trajectories generated by either TMD, SMD, or BMD differ. The $\overline{\text{rmsd}}_{ij}$ values calculated between unfolding trajectories generated with different perturbation methods are equivalent to values obtained by comparing multiple runs by the same method and simulation length.

Inspecting all trajectories with the visualization program VMD[49] identified the global features shared by all members from the same cluster. These common features are shown in Figure 4 by two representative trajectories for unfolding from each cluster. In cluster A trajectories, the protein unfolds by first extending and unwinding H1 and H3 from the ends. Breaking of the H1−H2 and H2−H3 interhelical contacts follows as a result of the stretching at the ends by perturbing h. The protein then extends into a linear chain. In cluster B trajectories, the H1−H2 hairpin always opens up with a flipping of H1 at an early stage of unfolding. The opening up of the H2−H3 hairpin occurs later. In contrast, these two events happen in the opposite order in cluster C, where H3 flips to lose contact with H2 earlier than the separation of H1 and H2.

The stronger dependence of the transition path on the progress variable than the method is evident from a similarity shared by cluster B and C paths regardless of the big rmsd between them. In these two clusters, the opening up of one of the two hairpins (H1−H2 in B and H2−H3 in C) happens

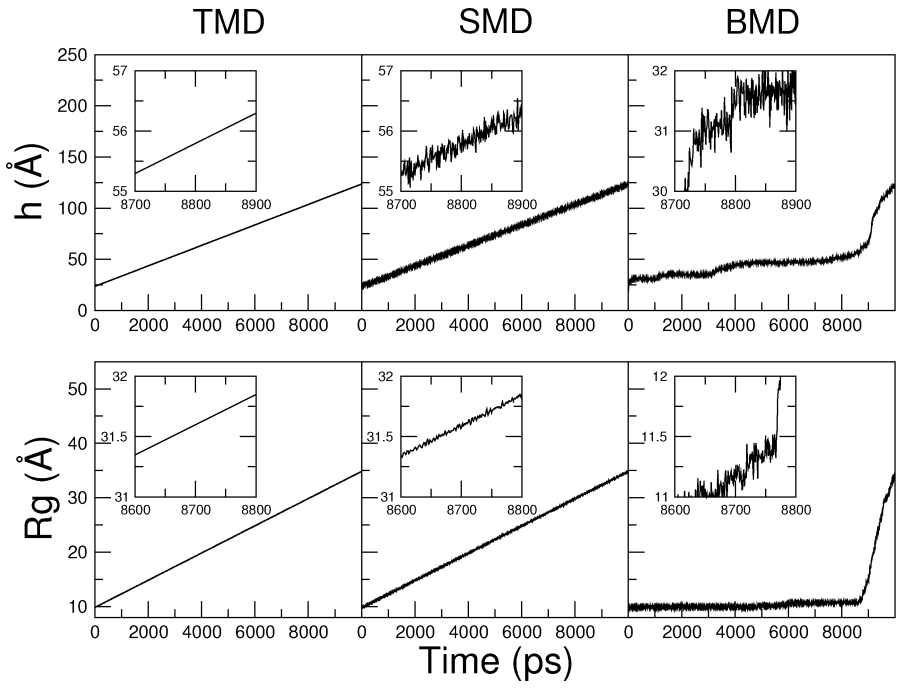


Figure 2. Progress variable as a function of time in representative BdpA unfolding trajectories generated by TMD, SMD, and BMD. The perturbed progress variable is the end-to-end distance (h) in the top panels and the radius of gyration (R_g) in the bottom panels. Insets show close-ups of a small portion of the progression curves.

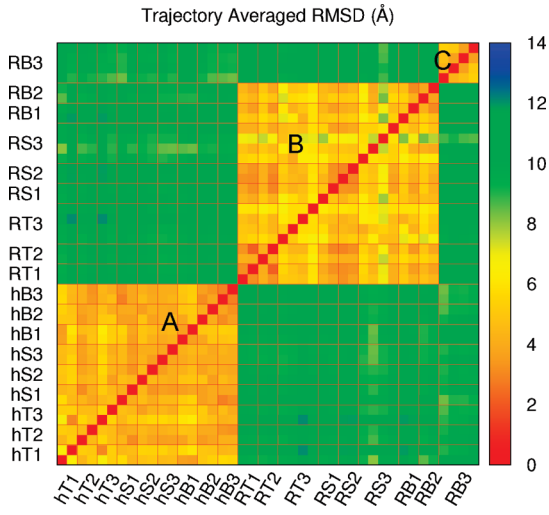


Figure 3. Pairwise trajectory averaged rmsd matrix for all BdpA unfolding trajectories. Each color square (pixel) shows the $\overline{\text{rmsd}}_{ij}$ value between a pair of unfolding trajectories. Trajectories are arranged in the same order on both axes so that the matrix is symmetric. Multiple trajectories with the same simulation condition are grouped together by thin lines in red, and the coding of simulation conditions is as in Table 1. Three clusters can be identified as follows: cluster A, all h-perturbed trajectories; cluster B, all R_g-perturbed trajectories except BMD, 10 ns ones; cluster C, R_g-perturbed, BMD, 10 ns trajectories.

as the first step of unfolding. This common feature is consistent with the external forces imposed by R_g perturbation in that the opening up of a hairpin will result in a big increase in R_g. In contrast, in all cluster A trajectories the unwinding of H1 and H3 from the ends precedes the breaking of the hairpin structure. This alternative pathway is also consistent with the pulling forces at the ends imposed by the end-to-end distance, h, perturbation.

Trajectories from the same cluster follow a common global trend but still display variability in structural details. One example of this variability is that H2 breaks in the middle (Figure 4, top right, $h = 60$ Å) and re-forms ($h = 76$ Å) before it is unwound in some but not all cluster A trajectories. Another example is that the unwinding of H1 accompanies H1−H2 separation in some but not all cluster B trajectories (Figure 4, middle left). Similar differences are also observed in cluster C trajectories (Figure 4, bottom). These local differences are not specific to a certain perturbation method. However, in general, the variability of R_g trajectories is greater than that of h trajectories in two aspects. First, R_g perturbation resulted in two distinct clusters (B and C) whereas h perturbation resulted in a single one (A). Second, R_g-perturbed trajectories have higher intracluster $\overline{\text{rmsd}}_{ij}$ values, as shown by more yellow-green elements in the similarity matrix in cluster B than in cluster A. This greater variability produced by the R_g bias is consistent with the more divergent nature of its corresponding forces.

BMD 10 ns R_g-Perturbed Trajectories Form a Distinct Cluster. Within R_g-perturbed trajectories, the RB3 trajectories generated by BMD in 10 ns (the longest time, with the smallest force constant) form a distinct group cluster C. As mentioned earlier, they differ from other R_g-perturbed trajectories in first opening up hairpin H2−H3 instead of H1−H2. In the three-helix-bundle structure of BdpA, both H1 and H3 form antiparallel interhelical tertiary interactions with H2, resulting in a nearly symmetrical topology. The interruption of one of these two tertiary interactions during unfolding can be considered a breaking of symmetry.[38] Assuming this symmetry to be perfect, its disruption should occur randomly at H1−H2 or H2−H3 as a result of the

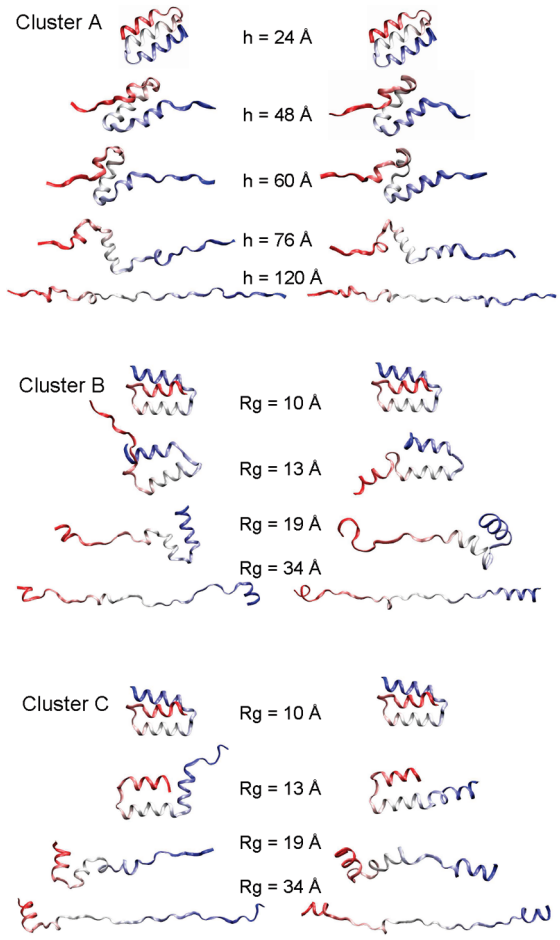


Figure 4. Snapshots unfolded to different extents for representative trajectories from each cluster. Helices H1, H2, and H3 are shown in red, white, and blue, respectively. These ribbon graphs show the sequence of key events involved in the unfolding, including the unwinding of individual helices and the interruption of different tertiary contacts. Two trajectories are shown for each cluster to exemplify variations within a cluster.

chaotic nature of molecular dynamics. However, here we observe H2−H3 breaking only in RB3 trajectories. This observation suggests an underlying picture in which both the protein's internal interactions and the external perturbation forces are slightly asymmetric and favor different symmetry-breaking positions, and the balance between the two asymmetric factors determines which contact breaks first. With short simulation lengths, the perturbation strength was high and all perturbation methods generated H1−H2 breaking trajectories only. As the simulation time increased, the perturbation strengths for all three methods were decreased, which would potentially cause a change of the balance between the external perturbation and internal interaction. Interestingly, in the longest simulation time (10 ns), only BMD sampled the alternative pathway, suggesting that BMD gives a smaller effective perturbation in the same simulation time. This smaller effective perturbation can be explained by the nonlinear sampling of BMD. With BMD, a larger portion of the total sampling time is spent before barrier-crossing events. Because barrier-crossing events are the critical events in determining which path to follow, BMD renders a weaker perturbation during these important events and thus sampled the alternative pathway in the longest simulation time.

Cluster C Has Lower PMF Values Than Cluster B. The TMD, SMD, and BMD perturbation methods were further examined for their ability to generate physically relevant transition paths by an assessment of the energetics of the unfolding trajectories. Because instantaneous energies are sensitive to local structural changes and contain a high level of noise, they were not used to compare the energetics of the trajectories. Instead, a potential of mean force (PMF) curve was calculated for a trajectory by using umbrella sampling and WHAM with snapshots taken from the trajectory as initial coordinates. The PMF curves aim to characterize the path obtained from trajectories generated for set conditions of perturbation method and length of simulation. It is recognized that there is variation among the curves because of limited sampling in the energy averaging on a rugged conformational landscape. Nonetheless, the PMF curves were found to be useful for comparative purposes.

The PMF with respect to R_g were calculated for all R_g-perturbed trajectories (clusters B and C) over the R_g range of 9.5−16.0 Å (see Methods for details). Figure 5 shows the average PMF and its variance as a function of R_g for each simulation condition. The average and variance were calculated from multiple (two to four) independent trajectories for each simulation condition and are shown in Figure 5a−c. Only the averaged PMF is shown in Figure 5d for clarity. All PMF curves share the same overall shape. There is a minimum at $R_g = 9.8$ Å, which is the R_g of the native structure. There are no other minima along the curves, a behavior also observed in BMD unfolding of the Fn3 domain.[6]

Paths generated by RB3 trajectories (cluster C) have the lowest PMF for R_g values > 11 Å, indicating an energetically more favorable unfolding path. Nevertheless, a comparison of shorter (0.5 and 2 ns) trajectories reveals that those generated by BMD are not more energetically favorable than SMD or TMD generated trajectories. Interestingly, the structural difference between RB3 and other R_g-perturbed trajectories for R_g > 11 Å is the breaking of H2−H3 contact or that of H1−H2. It is likely that the difference in PMF between RB3 and other trajectories is a reflection of the two different unfolding pathways described earlier. The fact that RB3 paths have lower PMF curves is also consistent with the earlier discussed notion that BMD renders a smaller effective perturbation in the same simulation time, thus samples an alternative, lower free energy pathway.

3.2. Lyn Activation/Deactivation. The second example system on which the three perturbation methods were tested is the conformational transition of the Lyn kinase CD. This process is a transition between two end states with well-defined structures, in contrast to the unfolding of BdpA.

The catalytic domain of Src-family kinases, such as Lyn CD, adopts different conformations in the active and inactive states shown in Figure 6. Major conformational changes of the CD upon activation are the opening of the cleft between the N- and C-terminal lobes (N-lobe and C-lobe), the rotation and translation of the αC helix, and rearrangement of the

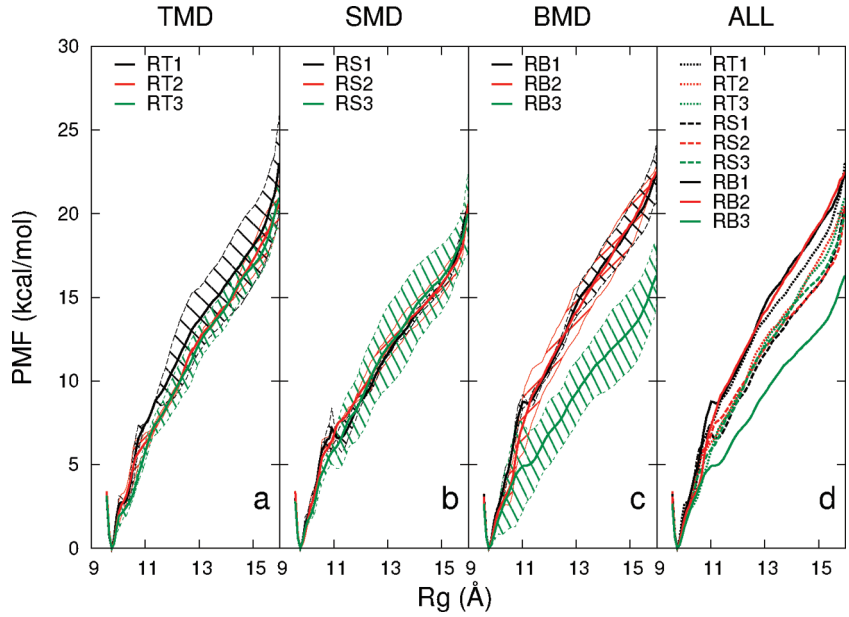


Figure 5. Potential of mean force (PMF) as a function of R_g for R_g-perturbed BdpA unfolding paths. The first three panels show the PMF curves, labeled by the codes shown in Table 1, for each time length for (a) TMD, (b) SMD, and (c) BMD. An average curve is calculated from PMF profiles of multiple trajectories (two to four) with the same simulation condition, and the shaded area shows the variance. (d) Average PMF curves.

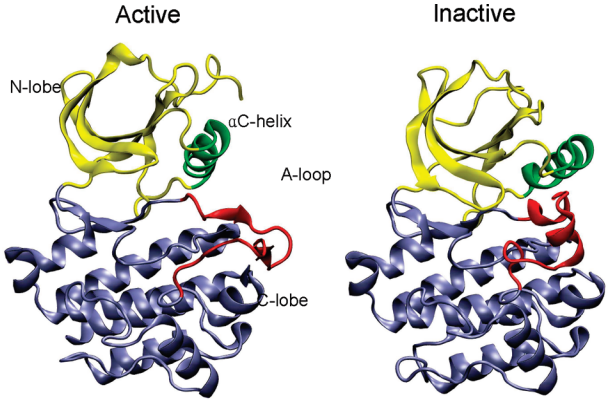


Figure 6. Comparison of the catalytic domain conformations of a Src kinase Lyn in the active (left) and inactive (right) states. The structures are from equilibrium molecular dynamics simulations[23] initiated with coordinates obtained by homology modeling with Lck[43] and Hck[44] crystal structures. Upon activation, the N-lobe (yellow) and C-lobe (blue) move apart, the helix C (green) moves and rotates inward, and the A-loop (red) deforms from two short helices to extended structure.

activation loop (A-loop). The 20-residue-long A-loop undergoes the most complex rearrangement. It adopts an extended conformation in the active state, whereas in the inactive state it folds back into the catalytic cleft, forming two short α helices.

Both activation and deactivation transition paths were calculated for the Lyn CD using the three perturbation MD methods. The two target structures were obtained by equilibrium MD with explicit water molecules initiated from homology modeled coordinates (see Methods). All three perturbation methods were applied with the MSID progress variable (see Methods for definition) to guide the transition to the target structure. Trajectories perturbing the root-mean-square deviation (rmsd) were also calculated with TMD. Details about the simulations are listed in Table 2.

Path Completeness. Unlike unfolding, the conformational transition process modeled in this example has a defined target configuration. By definition, TMD guarantees that the target configuration will be reached, while SMD only restrains the system to the target value with a defined force constant. With BMD, even the final reference point ρ_0 is not guaranteed to be the target value in a certain amount of time because it depends on spontaneous fluctuations of the system.

The final states of Lyn generated by the three methods are in accord with the above expectations. Figure 7 shows a superposition based on all heavy atoms of the target structure and the structure at the end of the perturbation MD run for representative deactivation simulations. TMD guided the transition to the exact target. With SMD, the final structure and the target configuration differ slightly (overall rmsd = 1.2 Å and A-loop rmsd = 2.8 Å), with differences in several backbone torsion angles in the C-terminal region of the A-loop. The final structure generated by BMD has the largest deviation (overall rmsd = 1.9 Å, A-loop rmsd = 3.4 Å) from the target of the three perturbation methods. The major deviation is again in the A-loop region but also involves differences at the secondary structure level. Specifically, the two short A-loop helices do not form.

Effect of Perturbation Methods and Progress Variable on Determining Path. Figure 8a shows the projection of all 20 trajectories onto two coordinates: Δrmsd of the A-loop and Δrmsd of the N-lobe. $\Delta\text{rmsd}_{\text{A-loop}}$ and $\Delta\text{rmsd}_{\text{N-lobe}}$ of a configuration are defined as the difference between the rmsd of this configuration to the active structure and the rmsd to the inactive structure for the heavy atoms of the A-loop (residues 406−526, c-Src numbering) and the N-lobe (resi-

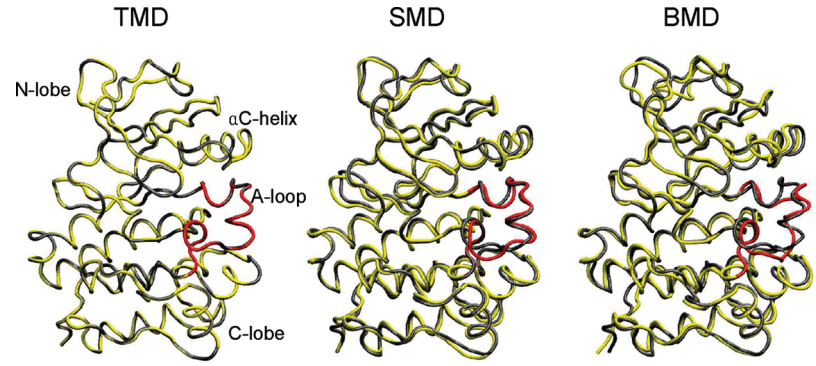


Figure 7. Final structures of Lyn CD deactivation trajectories generated by TMD, SMD, and BMD. These calculated final structures are shown in yellow except for the A-loop colored in red, whereas the native inactive structure, i.e., the target configuration, is shown in gray.

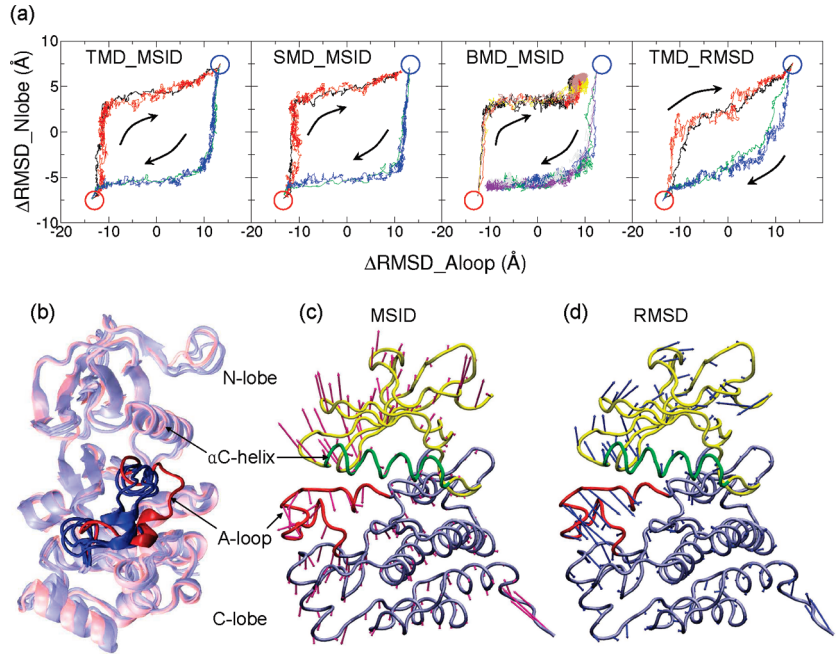


Figure 8. (a) Lyn CD activation/deactivation trajectories projected on the plane of Δrmsd of the N-lobe and Δrmsd of the A-loop. For each panel, multiple trajectories are shown in different colors. See Table 2 for color coding. Δrmsd = rmsd(to active) − rmsd(to inactive), superimposing the C-lobe of the protein. The active and inactive configurations are labeled by red and blue circles, respectively, and the transition directions are indicated by arrows. (b) Four snapshots with the N-lobe opened to the same extent from four Lyn CD activation trajectories generated by TMD−MSID, SMD−MSID, BMD−MSID and TMD−rmsd. For each trajectory, the first snapshot that reaches $\Delta rmsd_{N\text{-}lobe} = -5$ Å is shown. The three snapshots with MSID as the progress variable are colored in blue, whereas the one with MSID as the progress variable is colored in red. The A-loop region is shown in solid cartoon representation, while the rest of the protein is shown as transparent. (c, d) External forces acting on Lyn CD at the beginning of activation by perturbing MSID (c) or rmsd (d). The forces are averaged for each residue and scaled so that their average magnitudes are the same in both cases. rmsd induces relatively greater forces on the A-loop region compared to what MSID does.

dues 261−301 and 317−341), respectively, with the C-lobe superimposed.

Trajectories generated with the same progress variable, MSID, using three perturbation methods are similar on the two-dimensional Δrmsd surface, as shown in the first three panels of Figure 8a. These trajectories follow two different major paths depending on their direction being activation or deactivation. The two motions take place in two distinct steps that follow each other, as indicated by a sudden change in slope once one change is completed. For either direction, the lobe−lobe motion occurs before the rearrangement of the A-loop and the forward path is not the opposite of the reverse path. This behavior indicates that the energy barrier for the hinge-like lobe−lobe motion is lower than that for A-loop rearrangement which involves a change in secondary structure, and suggests that the global geometry of a transition path is dependent not only on the perturbation force, but also on the ease of the motions induced by the perturbation force. The observation of globally similar paths for all MSID-perturbed trajectories suggests that the dependence of the path on the perturbation method is minor, even though BMD generated deactivation paths that slightly differ from those generated by TMD and SMD at the end (Figure 8a, panel 3) due to a delayed motion of the αC helix.

The fourth combination of perturbation method and progress variable, TMD with the rmsd progress variable, generated paths globally different from all MSID paths. On the two-dimensional surface, these paths progress more linearly rather than stepwise, which suggests that the N-lobe and the A-loop move toward the target structure with a higher degree of cooperativity. This result is again in accord with the notion that the global path is largely determined by the choice of the progress variable.

The difference in the linearity between paths generated with MSID and rmsd as the progress variable can be seen structurally in snapshots from representative activation trajectories. Figure 8b shows an overlay of four snapshots from activation trajectories generated by the four combinations of perturbation method and progress variable. Upon activation, both the N-lobe and the A-loop open up from a closed conformation, and the four snapshots were chosen so that their N-lobes are opened to the same extent: $\Delta\text{rmsd}_{\text{N-lobe}} = -5$ Å. Figure 8b shows that, in the snapshot generated by perturbing rmsd (red), the A-loop motion has progressed further toward the active conformation while A-loops of the other snapshots generated by perturbing MSID (blue) remain close to the inactive state. It is worth noting that both MSID and rmsd are global measurements of the distance between two structures, and therefore as progress variables they are less distinct than h and R_g. Nonetheless, the subtle distinction does result in visible differences between generated paths, again showing the sensitivity of the perturbation MD methods to the choice of progress variable.

To examine the reason for the generation of different paths with the two progress variables, the forces imposed by perturbation on MSID or rmsd on each residue at the beginning of the activation are shown in Figure 8c,d. The magnitudes of the forces are scaled so that their average magnitude for the molecule is the same. It can be seen in the visualization that rmsd induces relatively greater initial forces on the A-loop region compared to the forces imposed by MSID. The variation and difference in direction in the force vectors explains the observed differences between the paths. Simulations perturbing the rmsd progress variable were not carried out with BMD and SMD, but results similar to that observed by TMD are expected.

4. Conclusions and Discussion

In the present study, we compared TMD, SMD, and BMD methods for guiding conformational transitions with external perturbation. Two distinct transition systems were explored: the unfolding of BdpA with implicit solvation and a heterogeneous end state, and the activation of Lyn CD in explicit water with two well-defined end states. Results from the three perturbation methods with two choices of progress variable and varying simulation length or biasing force constant showed that the perturbation methods generate similar transition paths for a given progress variable, as defined by global structural features. On the other hand, the path was strongly dependent on the choice of progress variable. Specifically, choosing to perturb R_g or h for unfolding BdpA determines whether the protein unfolds by stretching from the ends or by opening up one of the two interhelical contacts. Also, choosing to perturb MSID or rmsd to guide the conformational change of Lyn CD affects the cooperativity of the movements of its N-lobe and A-loop.

The path dependence on the progress variable was observed to be related to the direction and relative magnitude of the perturbation force, and these were the dominant influence regardless of the perturbation method. The alternative forces induce different motions of the molecule in certain directions and, together with the intrinsic ease of these motions, determine the transition path. In most cases, it is not clear that a priori knowledge can be used to decide which progress variable is a good reaction coordinate.

Although the three methods generated similar paths in terms of the global structural features for a given progress variable in most cases, the RB3 BdpA unfolding simulations generated by BMD with the lowest force constant identified an alternative class of paths, whereas longer SMD and TMD simulations generated similar paths to those generated by shorter simulations. The fact that this alternative class of paths is more energetically favorable is likely due to the softer and nonlinear perturbation rendered by BMD. BMD realizes a nonlinear updating scheme for moving along the progress variable, and thus the perturbation strength varies along a trajectory. In simulations of the same total length, BMD spends a greater amount of time going up a barrier than going down a slope. Thus the perturbation strength is comparatively low with BMD when the system searches for an easy route to overcome a free energy barrier. It should be noted that this nonlinear updating scheme of BMD would not guarantee the sampling of globally lower free energy paths, because an effectively lower perturbation strength only helps find an easy route to cross a local free energy barrier, which may or may not be a part of the global optimal path. In fact, the blindness to global features of the free energy surface is a common problem shared by all three perturbation methods.

In the conformational transition of Lyn CD, the backward paths are not the reverse copies of the forward paths with all three perturbation methods. This apparent violation of the microscopic reversibility suggests that the nonequilibrium results obtained with the perturbation MD methods do not approach the equilibrium results at the time lengths examined in our simulations. With the perturbation methods in the nonequilibrium scenario, the sequence of events along a path is determined by a combined effect of the perturbation and the internal ease of certain motions. Under the perturbation, the parts that feel the strongest perturbation and the weakest hindrance move toward the target first. Due to this lack of reversibility, caution should be taken when interpreting perturbation MD results to infer features such as sequence of events. Nevertheless, some insight into the ease of a certain motion can be obtained with these perturbation methods.

Global and/or local structural properties of the proteins should be considered when choosing a perturbation MD method. As our simulations on the conformational transition of Lyn CD show, BMD trajectories can be trapped in local minima in the presence of significant energy barriers (such as the A-loop rearrangement), preventing the system from

arriving at the target conformation. TMD and, for the most part, SMD ensure the completeness of the transition, although the likelihood of the path followed by TMD has not been assessed here. SMD closely resembles the AFM experiment, and it is preferred when direct analogy is desired. Further use of SMD allows the application of the Jarzynski equality to calculate equilibrium free energy differences.[50] Jarzynski's equality requires the external perturbation to be a component of the Hamiltonian as an explicit function of time, a condition satisfied by the perturbation form of SMD but by neither TMD nor BMD.[51]

Abbreviations and Symbols. BMD, biased molecular dynamics; TMD, targeted molecular dynamics; SMD, steered molecular dynamics; ρ, progress variable; h, end-to-end distance; R_g, radius of gyration; rmsd, root-mean-square deviation; MSID, mean square internal deviation.

Acknowledgment. This work was supported by National Institutes of Health (NIH) Grants GM039478 and GM083605 (C.B.P.), and a Purdue University reinvestment grant. E.O. and H.H. were supported by Purdue Research Foundation Fellowships. The authors acknowledge contributions from Bonnie Co in the initial stages of this work.

References

(1) Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. *Mol. Simul.* **1993**, *10*, 291–308.

(2) Schlitter, J.; Engels, M.; Kruger, P. *J. Mol. Graphics* **1994**, *12*, 84–89.

(3) Grubmueller, H.; Heymann, B.; Tavan, P. *Science* **1996**, *271*, 997–999.

(4) Leech, J.; Prins, J.; Hermans, J. *IEEE Comput. Sci. Eng.* **1996**, *3*, 38–45.

(5) Marchi, M.; Ballone, P. *J. Chem. Phys.* **1999**, *110*, 3697–3702.

(6) Paci, E.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 441–459.

(7) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568–1581.

(8) Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K. *Biophys. J.* **1998**, *75*, 662–671.

(9) Krammer, A.; Lu, H.; Isralewitz, B.; Schulten, K.; Vogel, V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1351–1356.

(10) Marszalek, P. E.; Lu, H.; Li, H. B.; Carrion-Vazquez, M.; Oberhauser, A. F.; Schulten, K.; Fernandez, J. M. *Nature (London)* **1999**, *402*, 100–103.

(11) Law, R. J.; Munson, K.; Sachs, G.; Lightstone, F. C. *Biophys. J.* **2008**, *95*, 2739–2749.

(12) Swift, R. V.; McCammon, J. A. *Biochemistry* **2008**, *47*, 4102–4111.

(13) Zou, J.; Wang, Y.-D.; Ma, F.-X.; Xiang, M.-L.; Shi, B.; Wei, Y.-Q.; Yang, S.-Y. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 323–332.

(14) Apostolakis, J.; Ferrara, P.; Caflisch, A. *J. Chem. Phys.* **1999**, *110*, 2099–2108.

(15) Bui, J. M.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15451–15456.

(16) Ferrara, P.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. B* **2000**, *104*, 4511–4518.

(17) Ma, J.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 11905–11910.

(18) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 252–260.

(19) Paci, E.; Caflisch, A.; Pluckthun, A.; Karplus, M. *J. Mol. Biol.* **2001**, *314*, 589–605.

(20) Morra, G.; Hodoscek, M.; Knapp, E. W. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 597–606.

(21) Li, Y.; Zhou, Z.; Post, C. B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7529–7534.

(22) Stultz, C. M. *Protein Sci.* **2006**, *15*, 2166–2177.

(23) Ozkirimli, E.; Post, C. B. *Protein Sci.* **2006**, *15*, 1051–1062.

(24) Levinson, N. M.; Kuchment, O.; Shen, K.; Young, M. A.; Koldobskiy, M.; Karplus, M.; Cole, P. A.; Kuriyan, J. *PLoS Biol.* **2006**, *4*, e144.

(25) Kastenholz, M. A.; Schwartz, T. U.; Hünenberger, P. H. *Biophys. J.* **2006**, *91*, 2976–2990.

(26) Perdih, A.; Kotnik, M.; Hodoscek, M.; Solmajer, T. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 243–254.

(27) Ozkirimli, E.; Yadav, S. S.; Miller, W. T.; Post, C. B. *Protein Sci.* **2008**, *17*, 1871–1880.

(28) Tikhonova, I. G.; Best, R. B.; Engel, S.; Gershengorn, M. C.; Hummer, G.; Costanzi, S. *J. Am. Chem. Soc.* **2008**, *130*, 10141–10149.

(29) Zhong, W.; Guo, W.; Ma, S. *FEBS Lett.* **2008**, *582*, 3320–3324.

(30) Matrai, J.; Jonckheer, A.; Joris, E.; Kruger, P.; Carpenter, E.; Tuszynski, J.; Maeyer, M. D.; Engelborghs, Y. *Eur. Biophys. J.* **2008**, *38*, 13–23.

(31) Carrion-Vazquez, M.; Li, H.; Lu, H.; Marszalek, P. E.; Oberhauser, A. F.; Fernandez, J. M. *Nat. Struct. Biol.* **2003**, *10*, 738–743.

(32) Sato, S.; Religa, T. L.; Fersht, A. R. *J. Mol. Biol.* **2006**, *360*, 850–864.

(33) Vu, D. M.; Myers, J. K.; Oas, T. G.; Dyer, R. B. *Biochemistry* **2004**, *43*, 3582–3589.

(34) Myers, J. K.; Oas, T. G. *Nat. Struct. Biol.* **2001**, *8*, 552–558.

(35) Guo, Z.; Brooks, C. L.; Boczko, E. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10161–10166.

(36) Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898–13903.

(37) Cheng, S.; Yang, Y.; Wang, W.; Liu, H. *J. Phys. Chem. B* **2005**, *109*, 23645–23654.

(38) Itoh, K.; Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7298–7303.

(39) Binnig, G.; Quate, C. F.; Gerber, C. *Phys. Rev. Lett.* **1986**, *56*, 930–933.

(40) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665–9672.

(41) Im, W.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

(42) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(43) Yamaguchi, H.; Hendrickson, W. A. *Nature (London)* **1996**, *384*, 484–489.

(44) Schindler, T.; Sicheri, F.; Pico, A.; Gazit, A.; Levitzki, A.; Kuriyan, J. *Mol. Cell* **1999**, *3*, 639–648.

(45) Paci, E.; Smith, L. J.; Dobson, C. M.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 329–347.

(46) Markwick, P. R. L.; Doltsinis, N. L.; Schlitter, J. *J. Chem. Phys.* **2007**, *126*, 045104.

(47) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.

(48) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

(49) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14* (33−8), 27−8.

(50) West, D. K.; Olmsted, P. D.; Paci, E. *J. Chem. Phys.* **2006**, *125*, 204909.

(51) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.

CT9000153

JCTC Journal of Chemical Theory and Computation

CHARMM Additive All-Atom Force Field for Acyclic Polyalcohols, Acyclic Carbohydrates, and Inositol

Elizabeth R. Hatcher, Olgun Guvench, and Alexander D. MacKerell, Jr.*

Department of Pharmaceutical Sciences, 20 Penn Street HSF II, University of Maryland, Baltimore, Maryland 21201

Received February 2, 2009

Abstract: Parametrization of the additive all-atom CHARMM force field for acyclic polyalcohols, acyclic carbohydrates, and inositol is conducted. Initial parameters were transferred from the alkanes and hexopyranose carbohydrates, with subsequent development and optimization of parameters unique to the molecules considered in this study. Using the model compounds acetone and acetaldehyde, nonbonded parameters for carbonyls were optimized targeting quantum mechanical interaction data for solute−water pairs and pure solvent thermodynamic data. Bond and angle parameters were adjusted by comparing optimized geometries to small molecule crystal survey data and by performing vibrational analyses on acetone, acetaldehyde, and glycerol. C−C−C−C, C−C−C−O, C−C−O−H, and O−C−C−O torsional parameters for polyol chains were fit to quantum mechanical dihedral potential-energy scans comprising over 1500 RIMP2/cc-pVTZ//MP2/6-31G(d) conformations using an automated Monte Carlo simulated annealing procedure. Comparison of computed condensed-phase data, including crystal lattice parameters and densities, NMR proton−proton couplings, densities, and diffusion coefficients of aqueous solutions, to experimental data validated the optimized parameters. Parameter development for these compounds proved particularly challenging because of the flexibility of the acyclic sugars and polyalcohols as well as the intramolecular hydrogen bonding between vicinal hydroxyls for all of the compounds. The newly optimized additive CHARMM force field parameters are anticipated to be of utility for atomic level of detail simulations of acyclic polyalcohols, acyclic carbohydrates, and inositol in solution.

Introduction

Glucose and related six-carbon monosaccharides exist in aqueous solution in equilibrium between the thermodynamically favored cyclic pyranose form and the linear aldose form. Reduction of the aldehyde functionality in the aldose form to an alcohol yields a linear polyalcohol or polyol, also commonly referred to as an "alditol" or "sugar alcohol." Such six-carbon sugar alcohols, as well as related cyclic and shorter-chain linear polyols (Figure 1), have both biological and industrial significance. One example is the conversion of the linear aldose sugar D-glucose to the alditol D-glucitol (also known as sorbitol) by reduction of the aldehyde group at C1 and subsequent oxidation at C2 to produce the linear ketose sugar D-fructose. Elevated conversion from glucose to sorbitol to fructose in humans is a factor in diabetes, and pharmacologic inhibition of the enzyme aldose reductase that catalyzes the reaction is a form of therapy.[1−3] In addition to participating in metabolism, these compounds also participate in cell signaling. Inositol 1,4,5-trisphosphate, a derivative of the cyclic polyol inositol (Figure 1), is produced from the hydrolysis of the membrane phospholipid phosphatidyl inositol 4,5-bisphosphate and acts as a secondary messenger that signals a rapid release of calcium from intracellular stores.[4,5] In addition to their biological significance, this class of compounds and their derivatives have industrial applications as sugar substitutes, surfactants, lubricants, and even explosives, motivating their structural, dynamic, and thermodynamic characterization.

Molecular dynamics (MD) simulations are a powerful and flexible way of studying structure, dynamics, and thermo-

* Corresponding author e-mail: alex@outerbanks.umaryland.edu.

10.1021/ct9000608 CCC: $40.75
Published on Web 04/27/2009

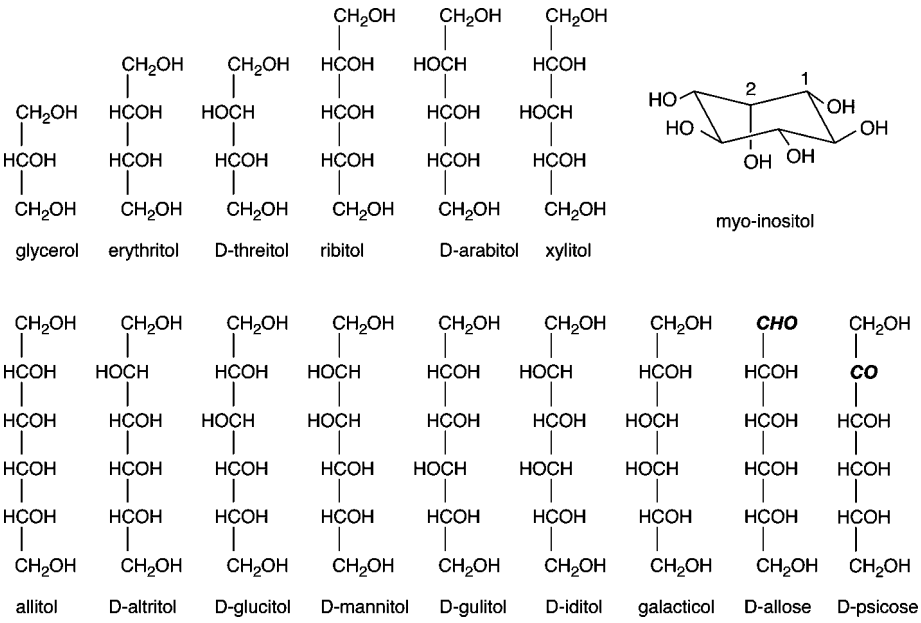


Figure 1. Acyclic polyalcohols, acyclic carbohydrates, and inositol. In the linear compounds, C1 is at the topmost position of the carbon chain; the aldehyde and ketone functionalities at C1 and C2 in D-allose and D-psicose, respectively, are in bold italics. In inositol, the C1 and C2 position are indicated with 1 and 2, respectively. Those compounds not designated with a D are meso compounds.

dynamics at an atomic level of detail. However, MD simulation results of a compound are only as reliable as the force field used to describe its structural and energetic properties. Accordingly, there has been much effort toward improving carbohydrate force fields, with focus primarily on cyclic pyranoses such as glucose.[6−13] Owing to the vast array of carbohydrates and their derivatives found in both biological and industrial settings, a comprehensive carbohydrate force field remains to be fully developed and validated. To maximize its utility, such a carbohydrate force field should also be compatible with available protein, lipid, and nucleic acid force fields since carbohydrates rarely participate in biological processes without interaction with these other three major biomolecular classes.

The present work describes the development of CHARMM force field parameters for linear polyalcohols, inositol, and linear sugars. These parameters have been developed for compatibility with the existing protein,[14,15] lipid,[16,17] and nucleic acid[18,19] CHARMM all-atom additive force fields[20,21] and extend the library of available carbohydrates, which had been previously limited to cyclic hexopyranoses.[22] Bond and angle internal parameters were transferred from previous work[22] with some modifications based on a survey of the Cambridge Crystallographic Database (CSD)[23] and vibrational analyses. Torsional parameters were fit to the relaxed quantum mechanical (QM) potential-energy surfaces at the RIMP2/cc-pVTZ//MP2/6-31G(d) level using an automated Monte Carlo simulated annealing procedure.[24] The nonbonded parameters for the linear polyols and inositol were transferred from the hexopyranoses,[22] whereas the nonbonded parameters for the carbonyl in aldoses and ketoses were fit to reproduce heats of vaporization and molecular volumes for neat liquids of representative model compounds as part of the present work. Parameter validation involved pure solvent calculations to compare to experimental heats of vaporization and molecular volumes, crystal simulations to compare to crystal lattice parameters, and aqueous-phase simulations to compare to experimental solution densities, diffusion constants, and NMR J coupling constants. Also addressed are the difficulties encountered during the parametrization process due to the flexibility of these compounds and the extensive intramolecular hydrogen bonding between the vicinal hydroxyl groups.

Methods

All molecular mechanics calculations were performed with the CHARMM program.[20,25] The all-atom additive CHARMM force field uses the energy potential $U(r)$ given in eq 1.

$$U(r) = \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{Urey-Bradley}} K_{UB}(S-S_0)^2 + \sum_{\text{dihedrals}} K_\chi(1+\cos(n\chi-\sigma)) + \sum_{\text{impropers}} K_{imp}(\phi-\phi_0) + \sum_{\text{nonbonded}} \varepsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{r_{ij}} \quad (1)$$

In eq 1, K_b, K_θ, K_{UB}, K_χ and K_{imp} are bond, valence angle, Urey−Bradley, dihedral angle, and improper dihedral angle force constants, respectively, while b, θ, S, χ, and φ are the bond distance, valence angle, Urey−Bradley 1,3-distance, dihedral angle, and improper dihedral angle, respectively, where the subscript zero represents the equilibrium value. In the dihedral potential-energy term, n is the multiplicity and σ is the phase angle as in a Fourier series. The nonbonded interaction energy between atoms i and j is separated into two terms, the Lennard−Jones (LJ) 6−12 term and the Coulomb term. For the nonbonded terms, ε_{ij} is the

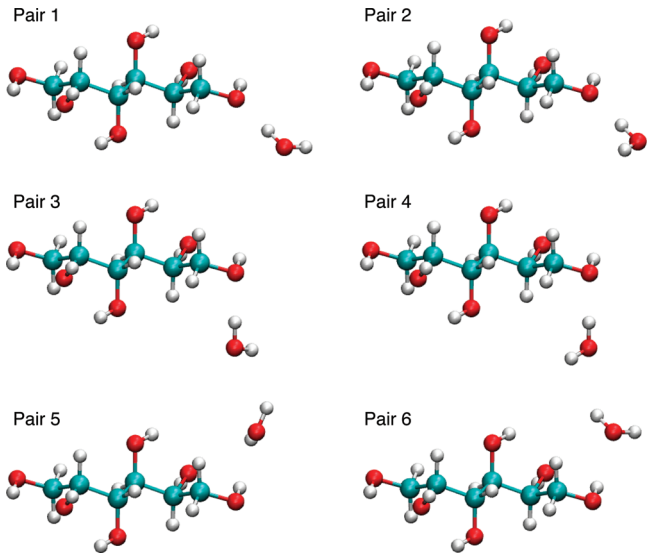


Figure 2. Allitol−water interaction orientations used for the water interaction calculations. VMD[70] is used to prepare the molecular graphics.

LJ well depth, $R_{\min,ij}$ is the 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. For the LJ parameters, the Lorentz−Berthelot combining rules are applied.[26]

Hydrogen-bonding water−solute pair interaction energies and distances were calculated using the standard additive CHARMM force field protocol so as to maintain compatibility with existing CHARMM biomolecular force fields.[20,21] Solute geometries were obtained from optimization at the MP2/6-31G(d) level of the conformation in the respective crystals obtained from the Cambridge Structural Database.[23] Using the optimized geometry of the monomer, water−monomer pairs were constructed, with the water internal geometry identical to that of the TIP3P water model.[27] Examples of these pair interactions are shown in Figure 2, where the water molecule is interacting with the terminal hydroxyl of allitol. In pairs 1, 2, 3, and 4 the hydroxyl oxygen is the hydrogen-bond donor, and in 5 and 6 the hydroxyl hydrogen is the hydrogen-bond acceptor. For pairs 1, 2, 3, and 4 the hydrogen of the water molecule is directed at the COH bisector. In pairs 1 and 2 the water molecule lies in the COH plane, whereas in pairs 3 and 4 the water molecule lies at a 120° angle to the COH plane. Pairs 5 and 6 are different because the water molecule in these interaction pairs acts as the hydrogen-bond acceptor; therefore, the COH hydroxyl is directed along the water HOH bisector. For pair 5, the HOH plane of the water is at a 90° angle to the COH plane, and for pair 6 the HOH and COH atoms are coplanar.

Reference data for comparison of molecular mechanics (MM) interaction energies and distances were generated by geometry optimization of the interaction distances at the QM HF/6-31G(d) level for each of the water−solute pairs above, with all other degrees of freedom constrained. The QM data cannot be targeted directly but are instead empirically scaled to account for the fact that the MM force field needs to be able to account for many-body effects in the condensed-phase. The CHARMM additive force field empirical scaling rules are well established[15,28] and such that the MM target distance is the R_{QM} − 0.2 Å and the MM target pair interaction energy (denoted "E_{QM}") is given by the expression $1.16(E_{QM,pair} - E_{QM,solute} - E_{QM,water})$. The energy-scaling factor of 1.16 and the offset of the QM distance by 0.2 Å account for limitations in the potential-energy function, and in the QM level of theory, and these empirical corrections lead to good agreement with condensed-phase properties, as shown in previous work.[20,21]

All of the C−C−C−C, C−C−C−O, O−C−C−O, and C−C−O−H dihedral parameters are fit to relaxed QM potential-energy scans. The Gaussian03 package[29] is used to optimize geometries at the MP2/6-31G(d) level of theory followed by single-point calculations performed at the RIMP2/cc-pVTZ level with the QCHEM program.[30] This level of theory has previously been shown to be sufficiently accurate for a number of systems including carbohydrates.[22,31] The target dihedral is scanned at 15° intervals from −180° to 165°, with the exception of inositol, which is scanned from 15° to 135° due to the constrained nature of the ring. The dihedral parameters are then fit to the QM dihedral scans using an automated Monte Carlo simulated annealing (MCSA) method.[24] In the MCSA method the selected dihedral parameters are fit simultaneously to minimize the root mean squared error (*RMSE*)

$$RMSE = \sqrt{\frac{\sum_i w_i (E_i^{QM} - E_i^{MM} + c)^2}{\sum_i w_i}} \qquad (2)$$

where E_i^{QM} and E_i^{MM} are the QM and MM energies of conformation i, w_i is a weighting factor for conformation i, and c is a constant that aligns the QM and MM data to minimize the *RMSE*. All of the six-carbon ($n = 6$) alditols are used in the fitting procedure (Figure 1), and all of the five-carbon ($n = 5$) and four-carbon ($n = 4$) alditols are used as the test set for the parameter validation. With inositol, the C−C−C−O, O−C−C−O, and C−C−O−H dihedrals are transferred from the hexopyranose parameters and only the C−C−C−C is fit (independently from the $n = 6$ alditols). For the dihedral parameters in the aldehyde and ketone groups in the linear carbohydrates D-allose and D-psicose (Figure 1), only the torsions containing non-hydrogen atoms and including the carbonyl atoms are parametrized and the other torsional parameters are transferred from the $n = 6$ alditols.

Parameter optimization and validation of the parameters is performed via a number of condensed-phase MD simulations. A cubic box containing TIP3P water molecules[27,32] with periodic boundary conditions is used for all aqueous simulations. Particle Mesh Ewald (PME)[33] with a 12 Å real-space cutoff is used to treat the long-range Coulombic interactions, and a force-switched smoothing function[34] with a range of 10−12 Å is used for the Lennard−Jones interactions, with a long-range correction applied beyond the truncation distance.[26] The SHAKE algorithm[35] is used to constrain all hydrogen-atom bonds to their equilibrium lengths and to maintain rigid water geometries. For the constant pressure−constant temperature (*NPT*) simulations the Nosé-Hoover thermostat[36,37] is used to maintain the

temperature and the Langevin piston barostat[38] is used to maintain the pressure. A leapfrog integrator[39] is used with a 1 fs time step for all of the simulations.

Pure solvent simulations are performed with a periodic box of 125 solvent molecules. The box of solvent molecules is minimized and then equilibrated for 50 ps followed by five production runs performed for 1 ns. The heat of vaporization ΔH_{vap} is calculated from the pure solvent simulation using the relation

$$\Delta H_{vap} = \langle U \rangle_{monomer} - \frac{\langle U \rangle_{box}}{N} + RT \tag{3}$$

Here, $\langle U \rangle_{monomer}$ is the average potential energy of the monomers calculated from five individual gas-phase simulations of all 125 molecules, with each simulation run for 500 ps. The $\langle U \rangle_{box}$ term is the average potential energy of the periodic box. N is the number of molecules in the box, R is the universal gas constant for an ideal gas, and T is the temperature.

The free energy of aqueous solvation ΔG_{sol} is calculated from the difference in free energy of a molecule in aqueous solution compared to that in the gas phase. ΔG_{sol} is calculated from the sum of nonpolar ΔG_{np} and electrostatic ΔG_{elec} free energies.[40]

$$\Delta G_{sol} = \Delta G_{np} + \Delta G_{elec} \tag{4}$$

ΔG_{np} is the sum of the repulsive and dispersive contributions, which are calculated using the Weeks, Chandler, Anderson decomposition of the LJ potential.[41] The repulsion term in the LJ potential is treated using a soft-core potential.[42] In the aqueous phase, free-energy calculations are performed using 1 molecule centered in a water box of 250 TIP3P water molecules. The aqueous system at each window is equilibrated for 50 ps and then simulated for 200 ps in the *NPT* ensemble. In the gas phase, Langevin dynamics are used with an infinite nonbond cutoff.[26,43] Since the gas-phase energies converge much more quickly, the gas-phase system is equilibrated for 10 ps and the production run is simulated for 100 ps. The simulations are performed at a temperature of 298 K and a pressure of 1 atm, which is consistent with experiment. The free-energy calculations are analyzed using thermodynamic integration[44] and the weighted-histogram analysis method[45] (WHAM). Additional details for calculating the free energy have been described previously.[40,46] Unlike all other condensed-phase simulations in the present work, due to software limitations, the long-range pressure correction is not part of the MD protocol for the free-energy simulations. Thus, the long-range contribution (LRC) from the LJ potential to the free energy of solvation is calculated as the difference in LJ energy of the aqueous system with a nonbond cutoff of 12 Å and a nonbond cutoff of 30 Å. The LRC is calculated from a 5 ps *NPT* simulation trajectory of the molecules in solution using coordinates saved every 100 fs and averaged over all values.

From the pure solvent simulation trajectories, the self-diffusion coefficient D_{sim} incorporates a system-size-dependent finite-size correction developed by Yeh and Hummer.[47]

$$D_{sim} = D_{PBC} + \frac{k_B T \zeta}{6\pi\eta L} \tag{5}$$

In eq 5, D_{PBC} is the diffusion coefficient calculated from a simulation with periodic boundary conditions to which the correction term is added. ζ is a constant of 2.837297, k_B is the Boltzmann constant, T is the temperature, η is the viscosity, and L is the length of the cubic simulation box. D_{PBC} is calculated from the slope of the mean square displacement of the C1 atom of all solute molecules in the simulation box versus time.[26] For diffusion coefficients of polyols in aqueous solutions of TIP3P water, eq 5 is further modified to take into account the low viscosity of TIP3P water relative to experiment.

$$D_{sim} = \left[D_{PBC} + \frac{k_B T \zeta}{6\pi\eta L} \right] \times 0.375 \tag{6a}$$

$$\eta = \eta_{TIP3P}(1 + 2.5\phi) \tag{6b}$$

Here, the scaling factor of 0.375 is applied to correct for the underestimation of the viscosity of water by the TIP3P model. The scaling factor is calculated from η_{TIP3P}/η_w, where $\eta_{TIP3P} = 0.35$ cP and $\eta_w = 0.93$ cP, the experimental viscosity of water. Equation 6b is the viscosity of a solution with the presence of a solute estimated by the Einstein formula,[48] where η_{TIP3P} is the viscosity of TIP3P (0.35 cP) and ϕ is the volume fraction of the solute. The method for calculating the simulation diffusion coefficient for a polyol−water mixture is similar to that previously used for a system of polyethylene oxide and polyethylene glycol.[49]

Complete crystal unit cells, obtained from the Cambridge Structural Database,[23] are used as starting structures for crystal simulations, with periodic boundary conditions applied in accordance with the length and angle parameters of the respective crystals. Each crystal system is minimized initially to remove bad contacts and is then equilibrated for 100 ps. After equilibration, the simulation is run for 2 ns. For all of the polyol crystals, the reference temperature is set to room temperature, 298 K, the temperature at which the crystals were obtained, and constant pressure is maintained at 1 atm by allowing independent variation in the crystal cell length parameters.

For the aqueous-phase MD simulations, a box containing 1100 waters and the number of solute molecules based on the experimental concentration is set up and then minimized using harmonic restraints with a force constant of 1 (particle mass) kcal mol^{-1} Å^{-2} amu^{-1} on only the solute molecules. The system is equilibrated for 500 ps, and then the equilibrated conformation is used as the starting conformations for five different unrestrained 1 ns runs, using different initial velocities for each of the runs to achieve improved statistics. The reference pressure of the glucitol and mannitol systems is 1 atm, and the reference pressure of the galacitol, xylitol, erythritol, ribitol, glycerol, and *myo*-inositol system is 3.5 atm, in accordance with the experimental conditions. The density of each system is calculated using the following equations

$$\rho = \frac{N}{\langle V \rangle} \tag{7a}$$

$$N = \frac{(N_{water} + N_{solute})\langle MW \rangle}{N_A} \tag{7b}$$

Table 1. Comparison of Optimized Water Interaction Energies and O···H Distance by HF/6-31G(d) QM Calculations and CHARMM Force Field for the Pair Conformations Shown in Figure *1*

		E_{QM} (kcal/mol)	E_{MM} (kcal/mol)	$E_{MM} - E_{QM}$ (kcal/mol)	R_{QM} (Å)[a]	R_{MM} (Å)	$R_{MM} - R_{QM}$ (Å)
allitol	O1, pair 1	−5.84	−5.87	−0.03	1.84	1.87	0.03
	O1, pair 2	−5.52	−5.84	−0.32	1.85	1.87	0.02
	O1, pair 3	−5.81	−5.25	0.56	1.82	1.88	0.06
	O1, pair 4	−6.25	−5.80	0.45	1.83	1.88	0.05
	O2, pair 1	−2.81	−3.38	−0.57	1.85	1.87	0.02
	O2, pair 2	−3.21	−3.86	−0.65	1.87	1.88	0.01
	O2, pair 5	−8.65	−8.83	−0.18	1.73	1.80	0.07
	O2, pair 6	−8.51	−8.80	−0.29	1.74	1.80	0.06
altritol	pair 1	−5.48	−5.73	−0.25	1.85	1.87	0.02
	pair 2	−4.83	−5.21	−0.38	1.87	1.88	0.01
	pair 3	−5.43	−5.12	0.31	1.85	1.89	0.04
	pair 4	−4.76	−4.35	0.40	1.86	1.89	0.03
ribitol	pair 1	−2.34	−3.83	−1.49	1.93	1.90	−0.03
	pair 2	−2.09	−2.88	−0.79	2.13	1.98	−0.15
	pair 5	−7.31	−7.14	0.17	1.76	1.82	0.06
	pair 6	−7.04	−7.00	0.04	1.78	1.82	0.04
threitol	pair 1	−5.46	−5.63	−0.17	1.85	1.88	0.03
	pair 2	−5.04	−5.54	−0.50	1.87	1.87	0.00
	pair 3	−5.40	−4.99	0.41	1.84	1.89	0.05
	pair 4	−4.83	−4.65	0.18	1.87	1.90	0.03
myo-inositol	pair 1	−5.71	−6.13	−0.42	1.87	1.88	0.01
	pair 2	−5.78	−6.57	−0.79	1.86	1.87	0.01
	pair 3	−4.47	−3.48	0.99	2.71	2.91	0.20
	pair 4	−6.20	−6.19	0.01	2.51	2.48	−0.03
average				−0.14			0.03

[a] 0.2 has been subtracted from the R_{QM} values.

where $\langle V \rangle$ is the average volume calculated from all five runs. N_{water}, N_{solute}, and N_A are the number of water molecules, the number of polyol solute molecules, and Avogadro's number respectively. $\langle MW \rangle$ is the average molecular weight of the system. Equation 7a is also used to calculate the density for neat liquids; however, in this case N is simply the number of molecules in the periodic box.

The J coupling constants for glucitol and mannitol are also calculated from the aqueous simulations described above. However, the coupling constants for arabitol, ribitol, and xylitol are calculated from aqueous-phase simulations at 1 atm and a molality of 0.5 mol kg^{-1} using the same protocol for the aqueous simulations. The dihedral value for the proton−proton coupling is calculated every 1 ps for each of the production runs. Moreover, the dihedral value is calculated for each of the solute molecules in the respective systems; therefore, depending on the concentration of the simulation the amount of torsional data differs. The J coupling is then calculated from the dihedral values for each snapshot using the generalized Karplus equation[50]

$$J = 0.8 \cos \phi + 10.2 \cos^2 \phi \quad (8)$$

where ϕ is the H−X−X−H dihedral angle. Manipulation of the Karplus equation given in eq 8 allows the fraction of trans conformers to be calculated[50]

$$F_{trans} = (J_{obs} - 3.0)/(9.4 - 3.0) \quad (9)$$

In eq 9, J_{obs} is the observed coupling constant.

Results and Discussion

Parameter Optimization. All parameter optimization was done in a self-consistent manner so that when one parameter was changed in a molecule all other parameters were tested and reoptimized as necessary. The presented empirical force field data reflect the final set of self-consistently-optimized nonbonded and bonded parameters.

Nonbonded Parameters. *Polyols.* The nonbonded parameters for the aliphatic and hydroxyl moieties in the polyol compounds in Figure 1 were directly transferred from alkanes[51] and hexopyranoses,[22] and testing showed that no further optimization was required. This testing included the ability of the force field model to properly describe hydrogen bonding as compared to QM data, in terms of both hydrogen-bonding strength and distance.

Taken as a group, hydrogen bonds in the water-solute pairs are well represented using the transferred nonbonded parameters. The average error over all 24 interaction pairs examined (Table 1) is −0.14 kcal/mol for the interaction energy and 0.03 Å for the interaction distance, which shows a slight systematic underestimation of the energy and overestimation of the distance. The mean absolute error for the interaction energy and distance is calculated to be 0.43 kcal/mol and 0.04 Å, respectively. For allitol, both a terminal (O1) and nonterminal hydroxyl (O2) were investigated with good results for both types. Of particular note is that the interaction energies range from −2.81 to −8.65 kcal/mol in the scaled QM representation, and the MM representation faithfully captures this diversity in hydrogen-bond strength. MM water interactions for $n = 4$ (threitol), 5 (ribitol), and 6 (allitol and altritol) linear polyols as well as for inositol are all independently in good agreement with the QM results with the exception of a few weakly interacting pairs, i.e., pair interactions 1 and 2 for ribitol. However, in the cases where the weakly interacting pair has a large error, the strongly interacting pairs are in very good agreement. As the more favorable interactions dominate structural and

Table 2. Water Interactions for Acetone and Acetaldehyde[a] and Condensed-Phase Properties Including Heats of Vaporization, Molecular Volumes, Free Energies of Aqueous Solvation, and Self-Diffusion Coefficients for Neat Acetaldehyde, Acetone, and Glycerol

water interactions											
	dipole moments					interactions					
	X	Y	Z	total		E_{QM}	E_{MM}	$E_{MM} - E_{QM}$	R_{QM}[b]	R_{MM}	$R_{MM} - R_{QM}$
					acetone						
QM	0.00	2.77	0.00	2.77	pair 1	−5.26	−6.49	−1.23	1.89	1.69	−0.2
MM	0.00	3.57	0.00	3.57	pair 2	−7.09	−7.19	−0.10	1.82	1.71	−0.11
					pair 3	−5.44	−7.02	−1.58	1.90	1.69	−0.21
					acetaldehyde						
QM	2.38	1.10	0.00	2.62	pair 1	−4.76	−4.96	−0.20	1.92	1.83	−0.09
MM	3.28	0.64	0.00	3.35	pair 2	−6.62	−6.07	0.55	1.84	1.81	−0.03
					pair 3	−4.86	−5.50	−0.64	1.94	1.81	−0.13
					pair 4	−2.20	−2.50	−0.30	2.36	2.18	−0.18
					pair 5	−2.12	−2.48	−0.36	2.37	2.18	−0.19

condensed-phase properties								
	ΔH_{vap}	% error	V_m	% error	ΔG_{sol} (LRC)	absolute error	D_{sim} (cm^2/s)	% error
				acetone				
expt	7.41		123.00		−3.85		4.77E−5	
calc	7.37	−0.54	124.47	1.20	−5.02(−0.31)	−1.17	4.84E−5	1.5
				acetaldehyde				
expt	6.08		92.85		−3.50			
calc	6.21	2.14	94.84	2.14	−3.23(−0.20)	0.27		
				glycerol				
exp	21.90		121.40					
calc	18.80(22.34)[c]	−14.16(2.01)[c]	127.17	4.75				

[a] Acetone/acetaldehyde−water interaction pairs are shown in Figure S1 of the Supporting Information. [b] 0.2 has been subtracted from the R_{QM} values. [c] Value in parentheses calculated using the monomer energy based on glycerol conformations obtained from the pure solvent simulation; see text for details. Energies in kcal/mol and molecular volumes in Å^3.

dynamic properties, it is deemed more important to treat these interactions accurately. Of note is that all hydroxyls in all compounds have the same partial charges and LJ parameters assigned to them, which attests to their generality and transferability to similar compounds and provides evidence that an accurate force field can be developed with a relatively parsimonious set of nonbonded parameters.

As an additional test of the hydroxyl nonbonded parameters, the pure solvent properties of glycerol were determined. The results show an error of 4.8% in the molecular volume and −14.2% in the heat of vaporization (Table 2). These errors are significantly beyond the typical error for CHARMM pure solvents, especially considering that the pure solvent properties of ethylene glycol are in good agreement with experiment.[22] One possibility for the large error in the heat of vaporization is overly favorable intramolecular hydrogen bonding occurring between hydroxyl groups in the gas-phase simulations resulting from the need to over polarize hydroxyls to obtain correct scaled water-solute interaction energies appropriate for a condensed-phase additive force field. The resultant lowering of the value of $\langle U \rangle_{monomer}$ in eq 3 would lead to a larger heat of vaporization. To test this possibility, the gas-phase energy $\langle U \rangle_{monomer}$ was calculated using monomer trajectories extracted from the pure solvent simulation. In these trajectories it is expected that intramolecular hydrogen bonding would be diminished due to intermolecular hydrogen bonding of the monomers with surrounding molecules in the solvent environment. Using this approach to calculate $\langle U \rangle_{monomer}$ the resulting heat of vaporization is 22.34 kcal/mol, which is only a 2.0% error with respect to experiment. Thus, it appears that the significant

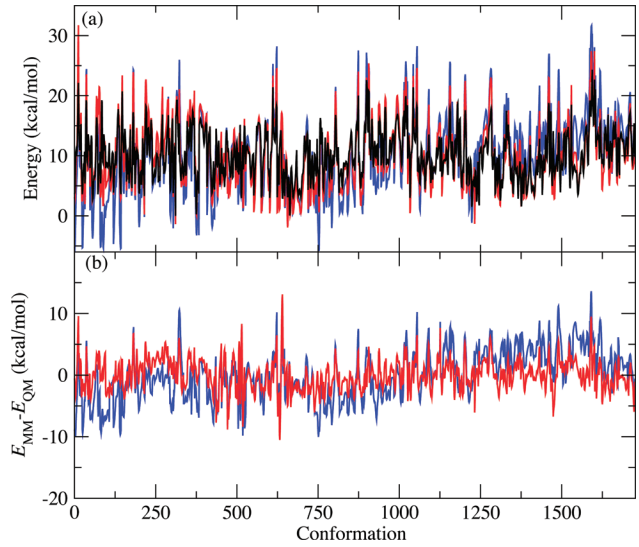


Figure 3. (a) QM and MM potential-energy scans for C−C−C−C, C−C−C−O, C−C−O−H, and O−C−C−O dihedrals for all $n = 6$ polyols (∼1730 conformations). QM scan is black; MM scan using optimized torsional parameters is red; MM scan using torsional parameters set to zero is blue. QM data have been offset using the global minimum as $E = 0$. MM scans have been root-mean square aligned with the QM scan (i.e., offset by the constant c given in eq 2). (b) $E_{MM} - E_{QM}$ using the optimized torsional parameters (red) and torsional parameters set to zero (blue).

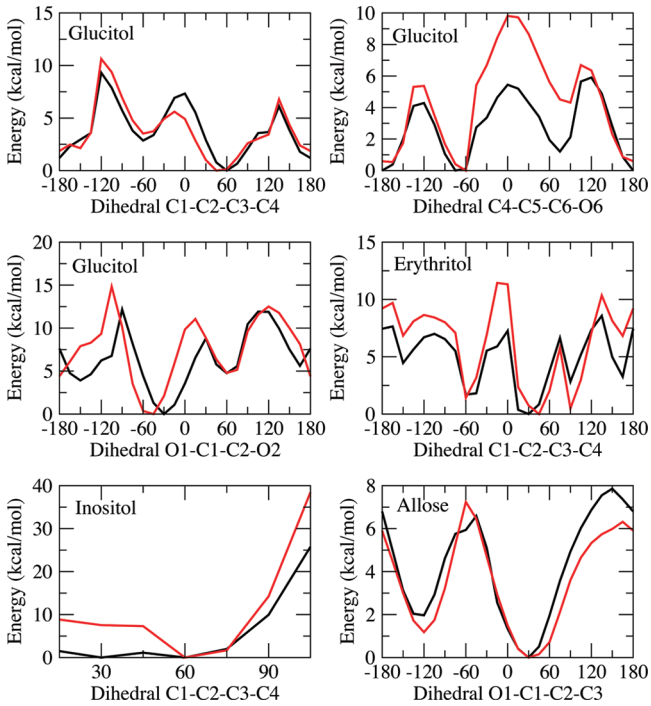


Figure 4. QM and MM potential-energy scans for dihedrals C1−C2−C3−C4, C4−C5−C6−O6, and O1−C1−C2−O2 in glucitol, C1−C2−C3−C4 dihedral in erythritol, C1−C2−C3−C4 dihedral in inositol, and O1−C1−C2−C3 dihedral in allose. QM results are black. The MM results, calculated using the optimized parameters, are red.

deviation in the heat of vaporization of neat glycerol is due to overestimation of intramolecular hydrogen bonding in the gas phase, although it should be noted that this effect will not influence the calculated molecular volume. Intramolecular hydrogen bonding in these compounds also complicates the parametrization of their conformational energetics, as discussed below.

Carbonyls. The nonbonded parameters for aldehyde and ketone carbonyls were explicitly optimized for application to the linear aldose and ketose forms of monosaccharides. Acetaldehyde and acetone were selected as model compounds for the parameter development process. Methyl parameters were as previously published,[51] and only the carbonyl C and O atoms and, with acetaldehyde, the aldehydic H LJ parameters and charges were optimized, with the charges of the adjacent bonded methyl carbons adjusted to yield a total molecular charge of zero. Target data for the parameter optimization were water−solute pair interaction energies and distances, dipole moments, heats of vaporization, and molecular volumes, while the free energies of aqueous solvation and the diffusion constant were calculated using the final optimized parameters.

Optimization of the carbonyl LJ parameters and partial charges yielded a set of parameters capable of reproducing all the target data. Overall, interaction energies with water are in satisfactory agreement with the QM target data, though in some cases the balance between the in-plane and the out-of-plane orientations is not ideal (Table 2 and Figure S1 of the Supporting Information). This problem is due to limitations in the form of the energy function and can, to a large extent, be corrected using an explicit representation of lone pairs.[52] However, to maintain consistency with the remainder of the CHARMM additive force fields, such an addition was not made. The final charges overestimate the QM MP2/6-31G(d) dipole moments by 28% and 29% for acetaldehyde and acetone, respectively. Such overestimation is required in the additive model to account for the implicit overpolarization required to accurately predict the condensed-phase properties.[21] The corresponding partial atomic charges along with the appropriate LJ parameters lead to excellent agreement for the pure solvent properties for both molecules, with the agreement within 2.2% of experiment[53−59] in all cases (Table 2).

Bonded Parameters. *Bonds and Angles.* All bond and angle parameters were initially transferred from similar existing CHARMM additive force field parameter values. Optimized geometries were compared to target data from a crystal database survey of the CSD. This analysis revealed systematic errors in geometries, in particular the bond lengths associated with C−C bonds, in which both carbons are hydroxylated. Accordingly, the respective geometric force field parameters were suitably optimized. A comparison of the optimized bond lengths and angles to the crystal survey data and QM data at the MP2/6-31G(d) level for compounds acetone and acetaldehyde is given in Table S2 in the Supporting Information. Vibrational analyses for acetone, acetaldehyde, and glycerol (Table S3 in the Supporting Information) were done both at the QM MP2/6-31G(d) level and in the MM representation, and the MOLVIB utility in CHARMM was used to assign the bond and angle contributions to internal normal modes as described by Pulay.[60] MM force constants were optimized as required to reproduce the scaled QM frequencies[61] to complete the bond and angle parameter optimization.

Dihedrals. Parameters associated with methyl rotations in acetone and acetaldehyde were readily optimized to yield excellent agreement between QM and MM methyl dihedral scans (O=C−C−H) (Supporting Information, Figure S4). In the case of the C−C−C−C, C−C−C−O, O−C−C−O, and C−C−O−H dihedrals, the torsional parameters were fit to QM RIMP2/cc-pVTZ//MP2/6-31G(d) potential-energy scans performed on linear polyols. These parameters were simultaneously fit to conformations with relative energies below a cutoff value of 15 kcal/mol for scan points on all of the $n = 6$ linear polyols, yielding approximately 1500 different conformational energies. Prior to fitting, with the parameters for the targeted dihedrals set to zero, the total *RMSE* (eq 2) for all of the conformations, including those above the energy cutoff of 15 kcal/mol (∼1730 different conformations), was 4.0 kcal/mol. Following fitting, which included the 1-, 2-, and 3-fold terms for each dihedral, sampling the force constant from 0 to 3 kcal/mol and sampling phase angles of either 0° or 180°, the *RMSE* was 2.5 kcal/mol. Using the dihedral parameters fit to the $n = 6$ sugar alcohols, the *RMSE* of the C−C−C−C dihedral for the $n = 5$ linear polyols is calculated to be 1.60 kcal/mol and for the $n = 4$ polyols to be 1.87 kcal/mol, as compared to values of 2.20 and 3.85 kcal/mol, respectively, with the targeted dihedral parameters set to zero. These data demonstrate both that the fitting proce-

Table 3. Crystal Lattice Parameter and Volumes Calculated from Crystal Simulations

compound	CSD ID	*R* factor	*A*	% error	*B*	% error	*C*	% error	volume	% error
					$n = 6$					
allitol	ALITOL01	0.04	4.71		13.41		6.62		411.16	
			4.96	5.35	13.81	3.03	6.82	3.14	445.30	8.30
altritol	JOJZOX	0.03	4.90		5.18		16.26		409.11	
			5.61	14.51	5.16	−0.39	16.11	−0.89	446.39	9.11
galacticol	GALACT	0.05	8.45		11.50		9.04		808.64	
			8.66	2.53	11.79	2.45	9.74	7.69	857.58	6.05
glucitol	GLUCIT01	0.07	8.68		9.31		9.73		785.86	
			8.71	0.44	9.48	1.84	10.11	3.89	834.77	6.22
mannitol	DMANTL07	0.03	8.69		16.90		5.55		815.40	
			8.96	3.07	17.22	1.91	5.61	1.03	865.23	6.11
					$n = 5$					
arabitol	ARABOL	0.04	9.21		4.86		15.49		692.85	
			9.36	1.59	4.95	2.00	15.39	−0.64	713.10	2.92
ribitol	RIBTOL	0.06	8.99		4.95		15.73		694.11	
			9.31	3.60	5.03	1.64	15.57	−1.04	720.91	3.86
xylitol	XYLTOL01	0.05	8.27		8.90		8.91		655.43	
			8.37	1.27	9.39	5.47	9.18	3.00	720.74	9.96
					$n = 4$					
threitol	PAGDEG	0.05	10.10		10.10		4.84		427.60	
			9.72	−3.66	9.72	−3.66	5.85	20.87	478.68	11.95
					$n = 3$					
glycerol	GLCROL	0.12	7.00		9.96		6.29		438.54	
			7.06	0.79	9.88	−0.78	6.82	8.43	474.91	8.30
					cyclic $n = 6$					
L-*chiro*-inositol	FOPKOK	0.03	6.87		9.13		6.22		373.68	
			6.93	0.93	9.44	3.41	6.33	1.85	402.09	7.60
average				2.77		1.54		4.30		7.31

Table 4. Calculated and Experimental[64,65] Densities at Different Concentrations of Polyols in a Box of Water with 1100 Water Molecules at $T = 298.15$ K and $P = 1$ or 3.5 atm

compound	molality (mol/kg)	N_{solute}	expt (g/cc)	calc (g/cc)	% error
mannitol ($P = 1$ atm)	0.1999	4	1.0092	1.0228	1.35
	0.5998	12	1.0318	1.0414	0.93
	0.8006	16	1.0426	1.0503	0.74
	0.9995	20	1.0517	1.0584	0.64
glucitol ($P = 1$ atm)	0.5085	10	1.0117	1.0368	2.48
	1.9508	39	1.0987	1.0921	−0.60
	4.0003	79	1.1606	1.1445	−1.39
	5.9945	119	1.2052	1.1793	−2.15
galacticol ($P = 3.5$ atm)	0.0698	2	1.0015	1.0179	1.64
	0.1492	3	1.0065	1.0202	1.36
xylitol ($P = 3.5$ atm)	0.1000	2	1.0022	1.0165	1.43
	1.0000	20	1.0427	1.0472	0.43
	2.7000	53	1.1019	1.0925	−0.85
erythritol ($P = 3.5$ atm)	0.4998	10	1.0142	1.0262	1.18
	1.0000	20	1.0298	1.0382	0.82
	3.0000	60	1.0806	1.0804	−0.02
ribitol ($P = 3.5$ atm)	0.1042	2	1.0022	1.0164	1.42
	0.5092	10	1.0208	1.0315	1.05
	3.1769	63	1.1121	1.1047	−0.67
glycerol ($P = 3.5$ atm)	0.5002	10	1.0075	1.0200	1.24
	1.0008	20	1.0171	1.0280	1.07
	3.0035	60	1.0497	1.0563	0.63
	4.9993	100	1.0750	1.0781	0.29
myo-inositol ($P = 3.5$ atm)	0.1000	2	1.0046	1.0191	1.44
	0.2480	5	1.0163	1.0288	1.23
	0.4994	10	1.0351	1.0443	0.89
average					0.64

dure leads to significant improvements in the conformational energetics of the targeted $n = 6$ compounds and also the transferability of the dihedral parameters to the shorter chain linear polyols.

Figure 3 shows the improvement in the MM conformational energies for the $n = 6$ target compounds. The change is visually apparent in the relative energies (Figure 3a) and especially so in the difference in QM and MM energies for both the parametrized and unparametrized energies (Figure 3b). It should be noted that all the relative energies for all the compounds are offset to a constant c, from eq 2, minimizing the total *RMSE* value. Figure 4 shows the QM and MM dihedral scans, using the optimized torsional parameters for a few representative compounds, and highlights the quality of the conformational energy fitting results. In all cases, the location of the minimum using the optimized dihedral parameters reproduces the QM results to within 15°. There is, however, a general trend toward overestimation of the energy barriers by the MM model. Two factors contribute to this; one is the fact that the targeted dihedral parameters are fit simultaneously to all of the $n = 6$ compounds, as required to maximize their generality and transferability (as evidenced by their applicability to the shorter polyols). The second is due to overestimation of intramolecular hydrogen bonding, again arising from the over-polarized hydroxyl groups that are required for proper condensed-phase behavior in an effective pairwise additive force field; these intramolecular hydrogen-bonding interactions are broken in the region of the energy barriers, exaggerating their energy differences relative to the minima.

Less extensive dihedral parameter optimization was required for the remaining compounds in Figure 1. The cyclic inositol was treated separately from the linear polyols, with parametrization of only the C−C−C−C dihedral, with the other dihedral parameters transferred from the hexopyranose force field. Using the MCSA method to fit the ring dihedral

Table 5. Diffusion Coefficients for Binary Water−Alditols Solutions

	molality (mol/kg)	volume fraction	size correction (cm^2/s)	expt D_{sim} (cm^2/s)[67,69]	calc D_{sim} (cm^2/s)[a]	% diff
glycerol	0.50	0.02	5.24E-6	1.10E-5	4.51E-6 (3.15E-6)	−59.04
galacticol	0.15	0.00	5.48E-6	6.36E-6	3.68E-6 (1.85E-6)	−42.08
glucitol	0.50	0.05	4.94E-6	6.58E-6	3.12E-6 (9.12E-7)	−52.62
mannitol	0.60	0.06	4.82E-6	6.67E-6	3.47E-6 (1.40E-6)	−47.90
myo-inositol	0.50	0.04	5.03E-6	6.08E-6	3.61E-6 (1.41E-6)	−40.58
average						−48.45

[a] The standard deviation of the calculated D_{sim} for the five production runs is given in parentheses.

Table 6. Calculated *J* Coupling Constants for Alditols Using the Karplus and the Fraction of Trans Conformers F_{trans}[a]

		coupling								
		$J_{(1,2)}$	$J_{(1',2)}$	$J_{(2,3)}$	$J_{(3,4)}$	$J_{(4,5)}$	$J_{(4',5)}$	$J_{(5,6)}$	$J_{(5',6)}$	avg diff
arabitol	expt	5.05	7.57	2.07	8.45	3.03	6.44			
	calc	1.56	7.44	2.50	4.30	2.61	4.24			
	diff	−3.51	−0.13	0.43	−4.15	−0.42	−2.20			−1.66
ribitol	expt	3.07	7.16	6.27						
	calc	1.09	8.05	3.54						
	diff	−1.98	0.89	−2.73						−1.27
xylitol	expt	4.34	6.90	4.43						
	calc	1.46	7.53	2.48						
	diff	−2.88	0.63	−1.95						−1.40
glucitol	expt	3.55	6.55	6.00	1.70	8.25		2.95	6.30	
	calc	2.84	5.90	2.75	2.79	2.73		3.08	5.49	
	diff	−0.71	−0.65	3.25	1.09	−5.52		0.13	−0.81	−1.39
mannitol	expt	2.94	6.43	8.99	1.02					
	calc	1.31	7.68	2.74	1.74					
	diff	−1.63	1.25	−6.25	0.72					−1.48
avg diff		−2.14	0.40	−2.75	−0.78	−3.36	−2.20	0.13	−0.81	

		fraction of trans conformers								
		F_{H1H2}	$F_{H1'H2}$	F_{H2H3}	F_{H3H4}	F_{H4H5}	$F_{H4H5'}$	F_{H5H6}	$F_{H5H6'}$	avg
arabitol	expt	32	71	0	85	0	54			
	calc	4	86	1	39	0	3			
	diff	−28	15	1	−46	0	−51			−18
ribitol	expt	0	65	51	51	0	65			
	calc	2	97	28	14	1	96			
	diff	2	32	−23	−37	1	31			1
xylitol	expt	20	60	23	23	20	60			
	calc	4	87	5	10	5	93			
	diff	−16	27	−18	−13	−15	33			0
glucitol	expt	20	56	37	0	73		5	51	
	calc	20	62	8	17	16		21	55	
	diff	0	5	−29	17	−57		16	4	−6
mannitol	expt	0	54	94	0	94		0	53	
	calc	2	90	18	2	19		4	91	
	diff	2	36	−75	2	−75		4	38	−10
avg diff		−8	23	−29	−15	−29	4	10	21	

[a] F_{trans} values are given in percentages.

in inositol, an *RMSE* value of 3.27 kcal/mol was obtained for an energy scan in which the ring was deformed from the favored chair conformation, an improvement on the *RMSE* of 4.23 kcal/mol before optimization. Additionally, for allose and psicose, only the dihedrals involving the carbonyl atoms and heavy atoms (C−C−C═O, O−C−C═O) required parametrization, since all others could be transferred from those for the linear polyols and acetone or acetaldehyde. The *RMSE* for the optimized allose and psicose dihedral parameters are calculated to be 0.75 and 0.69 kcal/mol as compared to 1.47 and 2.95 kcal/mol, respectively, for the unparametrized torsions. Though only allose and psicose were considered, the current set of parameters is applicable to all related ketoses and aldoses, owing to the demonstrated generality and transferability of the polyol parameters.

Parameter Validation. To validate the optimized parameters, pure solvent, crystal, and aqueous-phase MD simulations were performed to calculate both condensed-phase properties and conformational distributions and to compare them with experimental results. As well as testing the accuracy of the force field parameters, these calculations also gave insight into complications that arise during the parametrization process.

Acetone and Acetaldehyde. Validation of the aldehyde and ketone parameters was based on the calculation of the aqueous solvation free energy ΔG_{sol} and pure solvent diffusion coefficients for acetaldehyde and acetone. The calculated value for ΔG_{sol}, including the long-range LJ correction, is in excellent agreement with the experimental

value for acetaldehyde, with a difference of 0.3 kcal/mol (Table 2). The agreement for acetone is poorer, with the force field yielding a value that is 1.2 kcal/mol more favorable than the experimental value. While the level of agreement for acetone is not ideal, it is similar to that observed for a number of model compounds representative of amino acid side chains.[62,63] Furthermore, the diffusion coefficient for pure acetone is in excellent agreement and, combined with the water−solute interaction energy data (Table 2) indicates the model to have satisfactory energetic and dynamic properties.

Polyol Crystal Simulations. Crystal simulations were performed using crystals obtained from the CSD[23] to validate nonbonded parameters and conformational properties. The compounds included a number of tetritols ($n = 4$), pentitols ($n = 5$), and hexitols ($n = 6$) as well as glycerol and inositol. The selected crystals provide a thorough investigation of the crystal lattice parameters and are chosen based on diversity, purity, and resolution. To be consistent with experimental conditions, all of the crystals were simulated at room temperature (298 K).

The average sizes of the crystal unit cells from the simulations are systematically too large relative to the experimental values (Table 3). The average percent error of the simulated unit cell lengths *A*, *B*, and *C* for all of the selected polyols is 2.8%, 1.5%, and 4.3%, respectively, and the average unit cell volume is calculated to have an error of 7.3%. This overprediction in the average volume is consistent with what has been seen in previous studies for the hexopyranoses[22] and suggests limitations in the ability of a pairwise additive condensed-phase force field designed for liquid simulations to reproduce the crystal phase. The environments surrounding a molecule in the liquid versus the crystal state are considerably different, and the inability of the force field to quantitatively model the latter environment when parametrized to the former is not entirely surprising.

The heavy atom bonds, angles, and dihedrals are monitored during the simulations, and the averages reproduce the values in the crystal structure. This is particularly true for the bond lengths and angles because their equilibrium values were adjusted based on a crystal database survey. The differences between a selected set of experimental and simulated bond lengths, angles, and dihedrals are given in Table S5 of the Supporting Information. However, some of the dihedrals involving hydroxyl oxygens contrast to the values found in the crystal structure. In some cases, these differences cause errors in the calculated crystal lattice parameters, as hydrogen bonding that may occur in the crystal is not maintained during the simulation. For instance, in the altritol unit cell, a terminal hydroxyl hydrogen in one of the altritol monomers is hydrogen bonded to a hydroxyl oxygen in another altritol in an adjacent unit cell; this leads to the relatively large differences in the O1−C1−C2−C3 and O1−C1−C2−O2 dihedrals. During the crystal simulation, rotation about the torsions involving the hydrogens on these hydroxyls causes a loss in hydrogen-bonding capabilities. Subsequently, this causes an increase in unit cell length *A*, leading to larger errors in the unit cell *A* and volume of the unit cell.

Aqueous-Phase Simulations. To test the behavior of the polyols in aqueous solution, the environment in which it is anticipated the force field will primarily be applied, densities were calculated for molal solutions of glucitol, mannitol, ribitol, xylitol, galacticol, erythritol, glycerol, and *myo*-inositol varying in concetration from dilute (0.1 mol/kg) to highly concentrated (5 mol/kg) and compared to experimental values.[64,65] For consistency with experimental conditions, glucitol and mannitol were simulated at a temperature and pressure of 298 K and 1 atm, while all other compounds were simulated at 298 K and 3.5 atm. All of the calculated molecular densities reproduce experimental values within 3% error across the entire 50-fold difference in concentration and at ambient and elevated pressures (Table 4). Moreover, as the concentration increases there is a trend of decreasing error. The overall average error in volume, 0.64%, is much better than the overall crystal volume, emphasizing the applicability of the parameter set to heterogeneous liquid systems.

Polyol Diffusion Coefficients in Aqueous Solution. Diffusion coefficients for polyol−water solutions show consistent correlation with experimental values (Table 5). Complicating the analysis is the fact that TIP3P water has a self-diffusion coefficient larger than the experimental value, which motivated the addition of correction terms and scaling factors to the diffusion equation. The diffusion coefficients are computed using eq 6a, where the size-dependent correction term is calculated using the shear viscosity of TIP3P water model, which is 0.35 cP[66] (3.5E-4 kg/m/s), in conjunction with eq 6b for all polyol−water mixtures. The volume fraction $(\langle V\rangle_{\text{mixture}} - V_{\text{water box}})/V_{\text{water box}}$ is calculated from the average volume of the box for all five runs for the polyol aqueous-phase simulations. D_{PBC} is calculated at the given concentration for each of the five production runs and averaged. The computed results are compared to experimental results,[67] and the percent error is given in Table 5. The diffusion coefficients are systematically too low with a total average error of −48%. This error may be due to the fact that the D_{PBC} varies greatly between the five different production runs with a standard deviation of the average D_{PBC} for each polyol calculated to be ∼30%. The standard deviations are given in parentheses in Table 5. In addition, assumptions in the correction factors in eqs 6a and 6b may be limiting.

Alditol Conformational Sampling in Aqueous Solution. To investigate the ability of the model in treating the conformational properties of the alditols in solution, *J* coupling constants and conformer populations for all of the H^x-H^{x-1} hydrogens were computed from the aqueous-phase simulations of glucitol, mannitol, arabitol, ribitol, and xylitol. The dihedrals were computed for every solute molecule from snapshots taken every 1 ps from the MD simulations. Since the simulations of glucitol and mannitol are run with varying concentrations, the sampling of the dihedrals is much larger. For glucitol, the number of points sampled is ∼1.2 million, and for mannitol it is 260 000. The number of torsions calculated for arabitol, ribitol, and xylitol is 50 000. In the case of arabitol, ribitol, and xylitol, performing 5 more runs with different starting velocities and obtaining similar

dihedral probability distributions tested the convergence of sampling. *J* coupling values are calculated for each dihedral in each sample using the Karplus equation given in eq 8. Subsequently, the coupling constant is calculated as the average over all the *J* coupling values for each dihedral.

The H−H trans conformer probability was determined from the dihedral probability distributions, which were determined with a bin width of 3.6° between −180° and 180°. A trans conformer was considered to have any torsional value less than −135° and greater than 135°. The *J* coupling constants and the probability of H−H trans conformers F_{trans} are given in Table 6. The H−H trans conformer distributions for glucitol are shown in Figure S6 of the Supporting Information. As shown in Table 6, for all of the molecules studied, some of the *J* coupling values and F_{trans} values reproduce the experimental NMR data, although the differences are larger in some cases. For example, the average difference in F_{trans} involving one terminal hydrogen and one interior hydrogen (i.e., 1,2; 1′,2, 5,6, and 5′,6) is calculated to be 7 and the average difference involving two interior hydrogen atoms is −29. Therefore, F_{trans} are better described for those dihedrals involving the terminal hydrogen. However, it is important to note that the parameters for the associated dihedrals (terminal/nonterminal) have the same values in order to retain the transferability and simplicity of the parameters for these compounds. Therefore, the parameters overestimate the trans conformer population involving the terminal hydrogens and underestimate the trans conformer population involving only internal hydrogens. While such a compromise may represent an inherent limitation in the force field, limitations in the conversion from dihedrals to *J* coupling values may also contribute. Such limitations may be due to the use of a simplified generalized Karplus equation in this study (eqs 7a and 7b). Given the different environments of the terminal and nonterminal hydrogens, it may be more appropriate to use equations that are more detailed (e.g., that account for electronegativity of substituents) or optimized for the individual types of dihedrals. However, given the large number of compounds and classes of dihedrals, such a task is significant and beyond the scope of the present study.

Conclusions

The all-atom additive CHARMM force field for acyclic carbohydrates, acyclic polyalcohols, and inositol is presented. A full list of the topology and parameter information is included in Table S7 of the Supporting Information. The nonbonded parameters for the polyols are transferred from the hexopyranoses and tested using QM water−polyol interactions. The parameters for the carbonyl groups in linear aldoses and ketoses are optimized using the model compounds acetaldehyde and acetone and target QM data, model-compound−water interactions, and pure solvent properties. For all of the compounds in this study, the MM water interaction distances and energies are in good agreement with the scaled QM values. For the bond and angle parameters, initial optimization was performed using acetone, acetaldehyde, and glycerol, with the optimized parameters transferred to the larger acyclic carbohydrates. Concurrently, selected equilibrium bond distances were adjusted to fit CSD survey data. The dihedral parameters are optimized to fit an extensive set of QM dihedral energy scans on $n = 6$ linear polyols. The resulting parameters were then shown to be directly transferable to the $n = 4$ and 5 polyols. Validation of the optimized parameters was performed using condensed-phase simulations, including crystal simulations and aqueous-phase simulations. In the crystal simulations, all heavy atom bond, angle, and dihedral values reproduced the experimental crystal values. Crystal volumes calculated from the simulations are systematically too large by approximately 7%, consistent with a similar trend observed in crystals of hexopyanoses.[22] Although the crystal volumes are too large, molecular densities calculated from aqueous-phase simulations concentrations ranging from 0.1 to 5 mol/kg reproduce experimental data very well to within approximately 1%. As the parameters are anticipated to be used primarily for investigations of these compounds in aqueous solution, reproduction of the solution-phase results is deemed more important. Concerning the conformational properties in solution, trans conformer populations calculated using a generalized Karplus equation are overestimated for terminal dihedrals while being systematically underestimated for nonterminal dihedrals. Limitations in the applied Karplus equation may limit this analysis.

All of the compounds considered in the present study (Figure 1) are flexible molecules with multiple hydroxyls or carbonyl moieties that allow for extensive intramolecular hydrogen bonding. Such hydrogen bonding leads to difficulties in parametrizing a force field for these compounds. This is due to the need to overestimate the partial atomic charges in additive force field so as to account for the polarization of molecules that occurs in the condensed phase. Such increased charges, or overpolarization, is particulary problematic with these compounds as it leads to the extensive gas-phase intramolecular hydrogen bonding in the molecules being systematically overestimated; in the condensed phase it is assumed that competition between inter- and intramolecular hydrogen bonding yields a proper representation of interactions with the environment. Overestimation of intramolecular hydrogen bonding leads to several problems presented above. With glycerol, calculation of the heat of vaporization is confounded as the gas-phase intramolecular hydrogen bonding will tend to make the gas-phase energy too favorable, leading to the heat of vaporization being systematically too unfavorable. Calculation of the heat of vaporization using the gas-phase energy determined using the monomer conformations from the condensed-phase simulation (which is equivalent to determination of ΔH_{vap} based on only the intermolecular interactions in the condensed phase) yields much better agreement. This is consistent with previous results for ethylene glycol, where both the heat of vaporization and molecular volume were in good agreement with experiment.[22] However, it does not explain the molecular volume of glycerol being too large by 4.8%, which may be due to limitations in the nonbond parameters or the subtle balance in the competition of the inter- and intramolecular hydrogen bonding not being ideal. Another limitation due to overestimation of intramolecular hydrogen

bonding is its impact on the dihedral energy scans. This leads to the low-energy conformations, which maximize intramolecular hydrogen bonds, being artificially too favorable, which manifests itself in the MM energy surfaces overestimating the barrier height in the plots (Figure 4). This limitation leads to the *RMSE* for the energy surfaces to be higher than anticipated and may impact the resulting conformational properties of these molecules in solution, leading to poorer agreement with the NMR data. As this problem is due to inherent assumptions in the additive force field used in this study, it cannot be solved, though we attempted to alleviate this problem via targeting and validating the force field against a large body of target data. It is anticipated that polarizable force fields in which the need to overpolarize the gas-phase charge distribution is eliminated[68] may overcome these problems.

Acknowledgment. Financial support from the NIH (R01GM070855 (A.D.M.) and F32CA1197712 (O.G.)) and computational support from the Department of Defense High Performance Computing and the National Cancer Institute Advanced Biomedical Computing Center are acknowledged.

Supporting Information Available: Acetone and acetaldehyde−water interaction orientations; acetone and acetaldehyde geometric data; vibrational analysis for acetone, acetaldehyde, and glycerol; acetone and acetaldehyde dihedral scans; comparison of crystal and calculated bond lengths, angle, and dihedrals; conformer distributions of glucitol; topology and parameter files. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Costantino, L.; Rastelli, G.; Vescovini, K.; Cignarella, G.; Vianello, P.; DelCorso, A.; Cappiello, M.; Mura, U.; Barlocco, D. *J. Med. Chem.* **1996**, *39*, 4396.

(2) Singh, S. B.; Malamas, M. S.; Hohman, T. C.; Nilakantan, R.; Carper, D. A.; Kitchen, D. *J. Med. Chem.* **2000**, *43*, 1062.

(3) de la Fuente, J. A.; Manzanaro, S.; Martin, M. J.; de Quesada, T. G.; Reymundo, I.; Luengo, S. M.; Gago, F. *J. Med. Chem.* **2003**, *46*, 5208.

(4) Berridge, M. J.; Irvine, R. F. *Nature (London)* **1984**, *312*, 315.

(5) Volpe, P.; Salviati, G.; Di Virgilio, F.; Pozzan, T. *Nature (London)* **1985**, *316*, 347.

(6) Momany, F. A.; Willett, J. L. *Carbohydr. Res.* **2000**, *326*, 194.

(7) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541.

(8) Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236.

(9) Lii, J. H.; Chen, K. H.; Allinger, N. L. *J. Comput. Chem.* **2003**, *24*, 1504.

(10) Kony, D.; Damm, W.; Stoll, S.; van Gunsteren, W. F. *J. Comput. Chem.* **2002**, *23*, 1416.

(11) Lins, R. D.; Hünenberger, P. H. *J. Comput. Chem.* **2005**, *26*, 1400.

(12) Sixou, B.; Faivre, A.; David, L. G. V. *Mol. Phys.* **2001**, *99*, 1845.

(13) 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.

(14) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400.

(15) MacKerell, A. D., Jr.; 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.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(16) Feller, S. E.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2000**, *104*, 7510.

(17) Feller, S. E.; Gawrisch, K.; MacKerell, A. D., Jr. *J. Am. Chem. Soc.* **2002**, *124*, 318.

(18) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86.

(19) MacKerell, A. D., Jr.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105.

(20) MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. *CHARMM: The energy function and its paramerization 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, 1998; Vol. 1, p 271.

(21) MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584.

(22) 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.

(23) Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380.

(24) Guvench, O.; MacKerell, A. D., Jr. *J. Mol. Model.* **2008**, *14*, 667.

(25) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(26) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 1987.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(28) MacKerell, A. D., Jr.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 10559.

(29) 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.

(30) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(31) Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. A* **2006**, *110*, 9934.

(32) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. *J. Phys. Chem.* **1994**, *98*, 2198.

(33) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(34) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667.

(35) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(36) Nosé, S. *Mol. Phys.* **1984**, *52*, 255.

(37) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(38) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613.

(39) 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, p 136.

(40) Deng, Y. Q.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 8.

(41) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237.

(42) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025.

(43) Pastor, R. W. Techniques and Applications of Langevin Dynamics Simulations. In *The Molecular Dynamics of Liquid Crystals*; Luckhurst, G. R., Veracini, C. A., Eds.; Kluwer Academic Publishers: The Netherlands, 1994; p 85.

(44) Simonson, T. Free energy calculations. In *Computational biochemistry and biophysics*; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Watanabe, M., Eds.; Marcel Dekker, Inc.: New York, 2001; p 169.

(45) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.

(46) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2005**, *1*, 153.

(47) Yeh, I. C.; Hummer, G. *J. Phys. Chem. B* **2004**, *108*, 15873.

(48) Tanford, C. *Physical chemistry of macromolecules*; John Wiley and Sons: New York, 1961.

(49) Lee, H.; Venable, R. M.; MacKerell, A. D., Jr.; Pastor, R. W. *Biophys. J.* **2008**, *95*, 1590.

(50) Franks, F.; Dadok, J.; Ying, S.; Kay, R. L.; Grigera, J. R. *J. Chem. Soc., Faraday Trans.* **1991**, *87*, 579.

(51) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2005**, *109*, 18988.

(52) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.

(53) Vargaftik, N. B. *Handbook of Physical Properties of Liquids and Gases: Pure Substances and Mixtures*, 2nd ed.; Hemisphere: Bristol, PA, 1983.

(54) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257.

(55) Smith, B. D.; Srivastava, R. *Thermodynamic Data for Pure Compounds: Part A: Hydrocarbons and Ketones*; Elsevier: New York, 1986.

(56) McCall, D. W.; Douglass, D. C.; Anderson, E. W. *J. Chem. Phys.* **1959**, *31*, 1555.

(57) Ross, G. R.; Heideger, W. J. *J. Chem. Eng. Data* **1962**, *7*, 505.

(58) Smith, B. D.; Srivastava, R. *Thermodynamic Data for Pure Compounds: Part B: Halogenated Hydrocarbons and Alcohols*; Elsevier: New York, 1986.

(59) Stejskal, E. O.; Tanner, J. E. *J. Chem. Phys.* **1965**, *42*, 288.

(60) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. *J. Am. Chem. Soc.* **1979**, *101*, 2550.

(61) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502.

(62) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.

(63) Deng, Y. Q.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567.

(64) Hu, Y. F.; Zhang, Z. X.; Zhang, Y. H.; Fan, S. S.; Liang, D. Q. *J. Chem. Eng. Data* **2006**, *51*, 438.

(65) Blodgett, M. B.; Ziemer, S. P.; Brown, B. R.; Niederhauser, T. L.; Woolley, E. M. *J. Chem. Thermodyn.* **2007**, *39*, 627.

(66) Feller, S. E.; Pastor, R. W.; Rojnuckarin, A.; Bogusz, S.; Brooks, B. R. *J. Phys. Chem.* **1996**, *100*, 17011.

(67) Sartorio, R.; Wurzburger, S.; Guarino, G.; Borriello, G. *J. Solut. Chem.* **1986**, *15*, 1041.

(68) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory. Comput.* **2007**, *3*, 1927.

(69) *CRC Handbook of Chemistry and Physics*, 84th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 2003.

(70) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33.

CT9000608

JCTC Journal of Chemical Theory and Computation

Ab Initio Molecular Dynamics Simulations of the Infrared Spectra of $H_3O_2^-$ and $D_3O_2^-$

Martina Kaledin,* John M. Moffitt,† Craig R. Clark,‡ and Fareeha Rizvi

Chemistry and Biochemistry, Kennesaw State University, 1000 Chastain Rd., Box 1203, Kennesaw, Georgia 30144

Received November 24, 2008

Abstract: We present the infrared spectra of $H_3O_2^-$ and $D_3O_2^-$ calculated using MP2 direct molecular dynamics approach at temperatures of 100 and 300 K. The spectral peaks were assigned using the normal-mode analysis, instantaneous normal-mode analysis, isotopic substitution, polarized infrared absorptions, and analysis of the position−position correlation function. Our results predict the bridging hydrogen stretch between 600 and 900 cm^{-1} and bridging hydrogen bend vibrations between 1250 and 1650 cm^{-1}. We also examine two DFT methods (B3PW91 and B3LYP) and report on the differences between them and the MP2 spectra.

1. Introduction

Proton transfer in biological systems is thought to often proceed through hydrogen-bonded chains of water molecules.[1] Such chains could act as proton wires by providing an effective pathway for the rapid translocation of protons over long distances. A characterization of proton wires at the microscopic level is required for a better understanding of complex systems. The hydrated hydroxide ion, $H_3O_2^-$ is one of two fundamental structures (i.e., $H_3O_2^-$ and $H_5O_2^+$) involved in the proton transfer. The bridging proton in $H_3O_2^-$ and $H_5O_2^+$ is located between two oxygen atoms, e.g. $[HO-H-OH]^-$ and $[H_2O-H-OH_2]^+$. The structure and dynamics of $H_3O_2^-$ have been the subject of several theoretical studies.[2−18] The geometry of $H_3O_2^-$ has been predicted to be asymmetric with a longer H-bond of 1.342 Å and a shorter one of 1.125 Å.[11] In $H_3O_2^-$ two equivalent equilibrium structures of C_1 symmetry are separated by a transition structure of C_2 symmetry that comprises a low barrier[2,11] to proton transfer between two OH^- fragments. Parrinello et al.[3] investigated the quantum character of the shared proton in $H_3O_2^-$ and $H_5O_2^+$ system using ab initio techniques. These calculations at room temperature have shown that the shared proton in the strongly hydrogen bonded $H_5O_2^+$ behaved in an essentially classical manner, while in the $H_3O_2^-$ complex, quantum effects played a crucial role even at room temperature. Iyengar et al.[17] studied the properties of larger OH^- water cluster $OH^-(H_2O)_6$ that displays a hydroxide ion migration mechanism. Recently, Yang and Kuhn[18] investigated the effect of deuteration of the hydrated hydroxide ion using a quantum dynamical model and found zero-point energies in quantitative agreement with diffusion Monte Carlo calculations.

There has been progress in the experimental investigations of $H_3O_2^-$ and $H_5O_2^+$.[6−16,19−22] Price et al. reported the argon predissociation IR spectrum of $H_3O_2^-$ in the OH stretching region[19] and observed a strong sharp feature at 3653 cm^{-1}, which lies above vibrational band corresponding to the free hydroxide anion and below the region associated with the free OH stretch. Two weaker bands centered near 3100 and 3380 cm^{-1} were assigned to ionic bridging hydrogen vibration that is mixed with the H_2O bending mode. Another experimental study analyzed the hydration of the OH^- anion.[20] The vibrational spectra of cold hydrated ions $OH^-(H_2O)_{1-3}$ display a sharp feature in the OH stretching region, and in larger clusters a new feature appears in the region associated with interwater hydrogen bonding. A set of argon predissociation spectra for $H_3O_2^-$ and $H_5O_2^+$ were reported by Johnson's group.[21] Recent VSCF/CI[14] and

* Corresponding author phone: +1-770-423-6281; fax: +1-770-423-6744; e-mail: mkaledin@kennesaw.edu.

† Present address: University of Connecticut, Department of Molecular and Cell Biology, Storrs, CT 06269. E-mail: john.moffitt@uconn.edu.

‡ Present address: Georgia Institute of Technology, Department of Chemistry and Biochemistry, Atlanta, GA 30332. E-mail: craig.clark@gatech.edu.

10.1021/ct8004485 CCC: $40.75
Published on Web 03/31/2009

diffusion Monte Carlo[16,21] calculations for $H_3O_2^-$ indicate that the lowest-energy excitation involving displacement of the proton along the O−O axis would occur with much lower energies (650−720 cm^{-1}) than those predicted by standard normal-mode analysis at the global minimum.[2]

Motivated by uncertainties in theoretical and experimental predictions of bridging hydrogen vibrations, we recently reported the vibrational analysis of the $H_5O_2^+$ IR spectrum in the range 600 and 1900 cm^{-1}.[23] We presented calculations of the IR spectrum of $H_5O_2^+$ using the driven molecular dynamics (DMD) method and assigned the dominant spectral features near 1000 and 1780 cm^{-1} to the proton transfer coupled to the torsion motion and to the H_2O monomer bend coupled with proton transfer, respectively.

High level theoretical simulation of $H_3O_2^-$ IR spectra is a difficult task due to the high dimensionality of the problem and the floppy nature of the vibrational motion corresponding to proton transfer. Salahub et al.[8] tested DFT molecular dynamics simulations with nonlocal functionals of GGA and LAP families. They found out that the geometry, calculated vibrational spectra of the hydroxide complexes (especially the vibrations involving the hydrogen bonds), and the dynamical proton-transfer counting autocorrelation function are very sensitive to the choice of the functional. Previous theoretical studies of *N*-methylacetamide in solution[24] have shown that the B3PW91 functional together with the 6-31G(d,p) basis set provides a computationally efficient way to calculate vibrational frequencies and intensities in good agreement with experimental data for the systems involving hydrogen bonding.

In this work, we present a detailed study of the IR spectrum of $H_3O_2^-$ complex and its isotopomer $D_3O_2^-$ using the MP2 computational level with the direct molecular dynamics approach at two temperatures, 100 and 300 K. To our knowledge, this is the first attempt to calculate the IR spectrum of this system at that level of theory. A comparison of our theoretical simulations to the experimental observations has been made, and the spectral features have been assigned using the normal-mode analysis (NMA), instantaneous normal mode (INM) analysis, isotopic substitution, polarized infrared absorptions, and analysis of the position−position correlation function. We performed a standard NMA[25] by diagonalizing the Hessian matrix at the equilibrium structure and obtained nine normal-mode frequencies and corresponding normal mode vectors for $H_3O_2^-$ and $D_3O_2^-$. INM analysis was employed to predict the density of states and to attribute the spectral peaks to particular vibrational modes.

We also compared the results from MP2 and DFT (B3LYP, B3PW91) calculations to assess the accuracy and feasibility of the DFT approach to predict structure, vibrational frequencies, energetics, proton transfer barriers, proton population profile, and IR spectrum. The purpose of this work is to perform benchmark calculations using the direct molecular dynamics approach to obtain the IR spectra of $H_3O_2^-$ and $D_3O_2^-$. This approach will make it possible to analyze the vibrational motion of larger hydrated ions, for which an analytical form of the potential energy and dipole surfaces is not available due to high dimensionality.

2. Computational Methods

All calculations were carried out using the second-order Møllet−Plesset (MP2) perturbation theory,[26−28] with the basis set augmented by *s* and *p* diffusion functions on the oxygen atoms only (denoted as cc-pVTZ(aug-O:s,p), and B3LYP[29−31] and B3PW91[31,32] methods with the 6-31+G(d,p) basis set. The geometry of the $H_3O_2^-$ ground electronic state was optimized, and vibrational frequencies were obtained using the normal-mode analysis.

IR spectra of $H_3O_2^-$ and $D_3O_2^-$ were obtained by running direct molecular dynamics. The forces acting on the nuclei were computed from electronic structure calculations "on-the-fly" as the trajectory was generated at every time step. In this way, the construction of the explicit potential is avoided and the level of approximation in MD thus depends on the level of approximation for solving the Schrödinger equation, the choice of the method to calculate the potential energy and the basis set to describe the molecular orbitals. The direct molecular dynamics simulations were carried out on our Linux computer cluster using a shell script that interfaces the Gaussian 03 program[33] and our own suite of MD codes.

Spectra were collected for $J = 0$ using the expression for the intensity I

$$I(\omega) = \frac{Re}{\pi} \int_0^\infty dt\, e^{i\omega t} \langle \vec{\mu}(0) \cdot \vec{\mu}(t) \rangle_T \qquad (1)$$

where ω is frequency and $\langle \vec{\mu}(0) \cdot \vec{\mu}(t) \rangle_T$ is the dipole−dipole correlation function at temperature *T*. Trajectories were propagated at a constant temperature via Berendsen's thermostat.[34] The constant temperature was maintained with a 10 fs time response parameter.[34] The angular momentum was checked for conservation to ensure the accuracy of the propagator. The correlation function was then obtained as a time average along the trajectory.[35] Usually it is sufficient to run 10−20 long trajectories (order of picoseconds) to achieve convergence. For each temperature, the IR spectra were calculated from five trajectories propagated for 10 ps with a 0.5 fs time step (20 200 steps) at temperatures of 100 and 300 K using the B3LYP, B3PW91, and MP2 computational methods. We tested the convergence of the spectrum at the B3LYP level by running another five trajectories with different initial conditions. The shapes of the spectra from the two sets of simulations agree very well. The overlap of spectra from five and ten trajectories was 94% and 98% at 100 and 300 K, respectively. On the basis of this result, we concluded that five MP2 trajectories would yield a well-converged spectrum. The starting point for all trajectories was the equilibrium structure with randomly sampled velocities from the Boltzmann distribution. Initially the system was equilibrated for the first 200 steps. Then the IR spectrum was calculated by Fourier transform of dipole−dipole correlation function (eq 1), corrected with a quantum mechanical frequency-dependent factor.[36]

The orientation of $H_3O_2^-$ in the coordinate system and the atom labeling are shown in Figure 1. The *x*-axis is coincident with the OO-axis, the vibration of the shared proton between two oxygens. The IR spectra are plotted

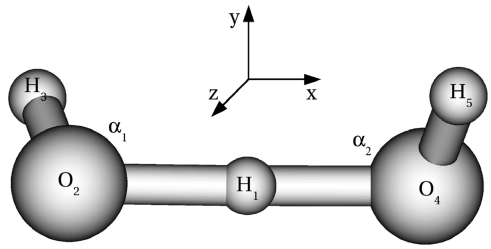


Figure 1. The orientation of $H_3O_2^-$ in the coordinate system and the atom labeling.

separately for the parallel and perpendicular components, $\langle \mu_x(0)\mu_x(t)\rangle_T$ and $\langle \mu_y(0)\mu_y(t) + \mu_z(0)\mu_z(t)\rangle_T$, respectively. This separation of polarization into the parallel (along *x*-axis) and perpendicular (along *y*- and *z*-axes) components facilitates the spectral assignment of the peaks based on the internal movements.

We have also evaluated the classical position−position correlation function[35] from the coordinates saved along MD trajectories. The position−position correlation function was separated into the parallel (*x*-axis) and perpendicular (*y*- and *z*-axes) components

$$C_{\parallel} = \langle x(0)x(t)\rangle_T$$
$$C_{\perp} = \langle y(0)y(t) + z(0)z(t)\rangle_T \quad (2)$$

and the spectrum was obtained by Fourier transform.

In order to better interpret our classical MD spectra and to assess quantum and anharmonic effects in the $H_3O_2^-$ system, an instantaneous normal-mode analysis[37−40] was performed. We characterized the spectrum in a certain frequency window by calculating the overlap[41] of the instantaneous normal mode vectors with the normal mode vectors of $H_3O_2^-$ at the minimum

$$O_{jk} = \sum_{i=1}^{3N} x_{ik}u_{ij} \quad (3)$$

where u_j is the *j*th normal mode vector, and x_k is the instantaneous normal mode vector. The contribution of each normal mode to the normal mode of the instantaneous structure can be defined as an average weight[23]

$$w_j = (1/N_f)\sum_{k=1}^{N_f} {O_{jk}}^2 \quad (4)$$

where N_f is the number of instantaneous structures in a given frequency window.

3. Results and Discussion

In this section, as a main result of this work, we present the IR spectra of $H_3O_2^-$ and $D_3O_2^-$ calculated using MP2/cc-pVTZ(aug-O:s,p) direct MD simulations. The geometry was optimized at the MP2 level of theory to a nonplanar structure with an asymmetric hydrogen bond. The structural parameters are listed in Table 1 and the labeling of the atoms is shown in Figure 1. The geometrical parameters for MP2/cc-pVTZ(aug-O: s,p) structure are in close agreement to those predicted by Xantheas[2] at the MP2/aug-cc-pVTZ level of theory. In our work, we have chosen truncated basis set cc-pVTZ(aug-O:s,p) to save computational time when runnning MP2 molecular dynamics simulations. Sauer and Dobler[42] calculated gas-phase IR spectra for hydrated water dimer $H_5O_2^+$ using MP2. They showed that the MP2/aug-cc-pVTZ and MP2/cc-pVTZ(aug-O) harmonic frequencies were very similar and concluded that the truncated basis set provides accurate O−H$^+$···O asymmetric stretch frequency.

Table 1. $H_3O_2^-$ Minimum Structure Using B3LYP, B3PW91, and MP2 Methods

internal coordinates[a]	B3LYP	B3PW91	MP2
R_1 (O_2−H_1), Å	1.309	1.222	1.331
R_2 (O_4−H_1), Å	1.154	1.222	1.130
R_3 (O_2−H_3), Å	0.966	0.963	0.962
R_4 (O_4−H_5), Å	0.965	0.963	0.960
α_1 (H_1−O_2−H_3), deg	106.5	104.5	103.5
α_2 (H_1−O_4−H_5), deg	103.8	104.5	101.7
δ_1 (O_2−H_1−O_4−H_5), deg	177.8	179.0	178.1
δ_2 (H_3−O_2−O_4−H_5), deg	110.5	111.0	106.8

[a] The labeling of the atoms is shown in the Figure 1.

Table 2. $H_3O_2^-$ and $D_3O_2^-$ Normal Mode Frequencies in cm^{-1} Using B3LYP, B3PW91, and MP2 Methods

	B3LYP		B3PW91		MP2	
	$H_3O_2^-$	$D_3O_2^-$	$H_3O_2^-$	$D_3O_2^-$	$H_3O_2^-$	$D_3O_2^-$
OH torsion	200	146	199	145	199	145
OO stretch	276	216	575	428	310	259
OH wag	515	416	225	160	519	404
OH rock	594	433	634	613	610	440
BH stretch, *x*-axis	967	790	631	465	1120	894
BH bend, *z*-axis	1446	1041	1489	1073	1433	1031
BH bend, *y*-axis	1672	1206	1668	1203	1686	1215
OH stretch	3825	2784	3872	2818	3864	2812
OH stretch	3844	2797	3873	2818	3890	2830

A. Dipole Spectra. The assignment of the spectral peaks was made by comparing our MP2 IR spectra and recent experimental IR spectra[16,21] to the previous high level calculations[14,15] and normal-mode frequencies listed in Table 2. Figure 2 illustrates the IR spectrum of the $H_3O_2^-$ calculated using MP2 at 100 and 300 K. The most dominant peak in our MP2 spectrum at 100 K (Figure 2a) centered at 250 cm^{-1} can be assigned to the OH torsional motion. Now we aim to investigate the spectral region associated with the motion of the shared proton in $H_3O_2^-$. The experimental argon predissociation spectrum (see Figure 2b in ref 21) exhibits a dominant spectral feature at 697 cm^{-1} and two small peaks around 1000 cm^{-1} assigned to bridging hydrogen (BH) motion (motion of a proton between the two oxygens) parallel and perpendicular to the hydrogen-bond axis, respectively. Another experimental study of the intramolecular bending transitions in the 1000−1900 cm^{-1} region[16] identifies the spectral feature near 1090 cm^{-1} as excitation of the shared proton along and/or perpendicular to the O−O axis[15] based on diffusion Monte Carlo[16] and the VSCF/CI methods.[14] Our MP2 spectrum in the region between 400 and 1800 cm^{-1} (Figure 2b) shows a broad feature between 500 and 1200 cm^{-1} with dominant peaks between 500 and 800 cm^{-1}, while the experimental spectrum[21] predicts a relatively narrow absorption peak at 697 cm^{-1} with Δ_{fwhm}

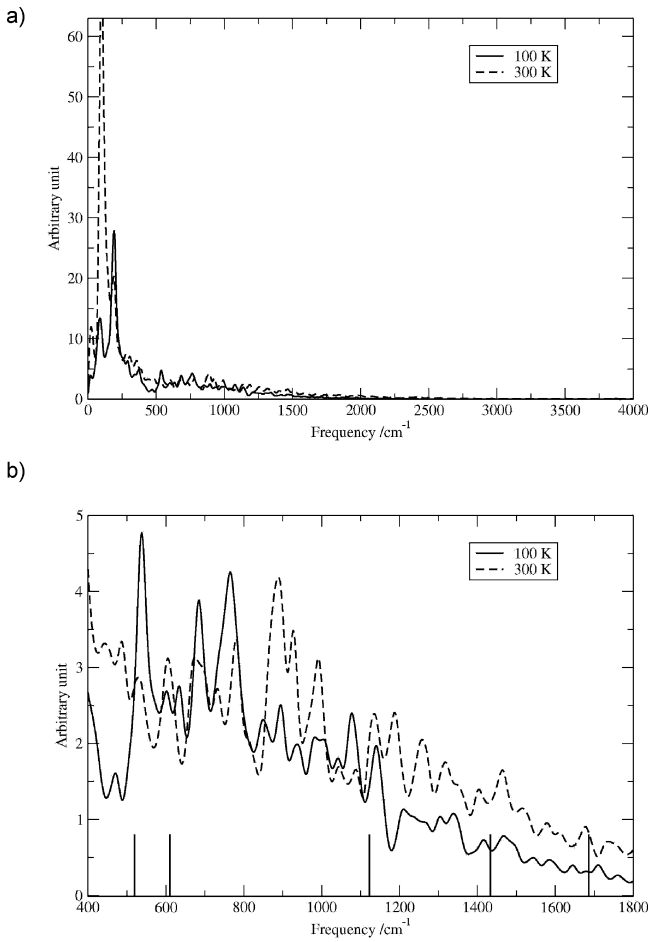


Figure 2. MD MP2 IR spectra for $H_3O_2^-$ at 100 and 300 K (a) in the full spectral range and (b) in the range from 400 to 1800 cm^{-1}. The harmonic frequencies are shown as sticks in the spectrum.

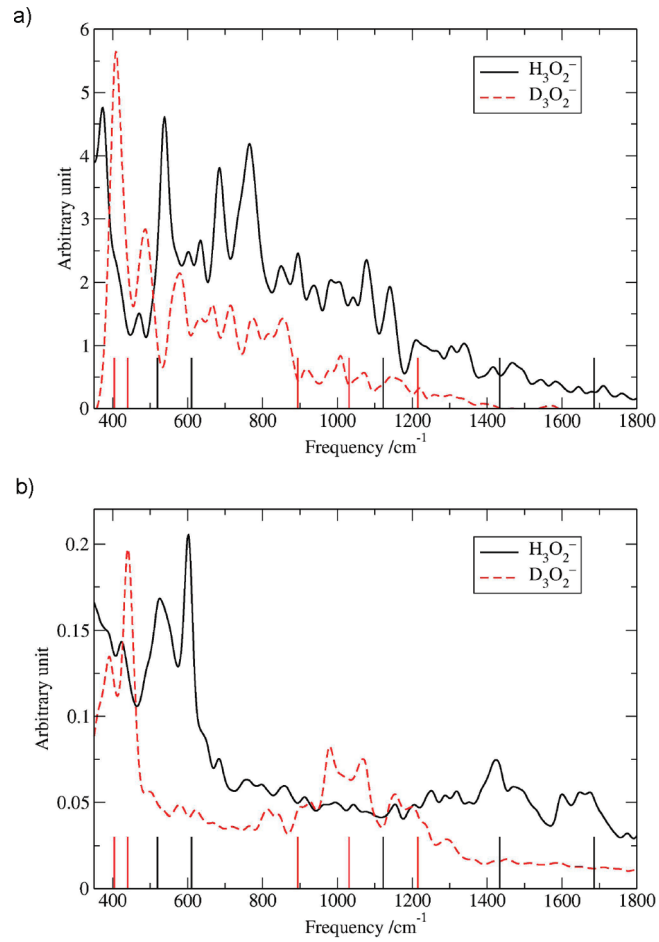


Figure 3. MD MP2 IR spectra at 100 K for $H_3O_2^-$ and $D_3O_2^-$ calculated using the dipole−dipole correlation function decomposed to (a) the parallel $\langle \mu_x(0)\mu_x(t)\rangle_T$ and (b) perpendicular $\langle \mu_y(0)\mu_y(t) + \mu_z(0)\mu_z(t)\rangle_T$ components. The harmonic frequencies are shown as sticks in the spectrum.

= 21 cm^{-1}. Simulations at 300 K show that this spectral feature broadens significantly.

The quantum studies of the in $H_3O_2^-$ and $D_3O_2^-$ vibrations[14,15] assigned the spectral bands between 500 and 800 cm^{-1} to the OO stretch, OH wagging, OH rocking, and BH stretch modes. Normal mode analysis predicts large x displacements for these vibrations. BH stretch parallel to the hydrogen-bond axis (x-axis in our case) is the most effective for modulation of the molecular dipole; therefore, it should have the stronger intensity in the spectrum. Our $H_3O_2^-$ and $D_3O_2^-$ IR spectra calculated at the MP2 level at 100 K (Figure 3) show that the parallel components of the dipole−dipole correlation function contribute significantly to the total spectral intensity, while contributions from the perpendicular components are much smaller. In the spectral range between 500 and 800 cm^{-1}, the MP2 IR spectrum of $H_3O_2^-$ calculated with only the parallel component of the dipole−dipole correlation function (Figure 3a) shows three infrared bands with similar intensities at 540, 685, and 765 cm^{-1} and two small peaks at 1075 and 1140 cm^{-1}. When the IR spectrum is calculated only with perpendicular components of the dipole−dipole correlation function (Figure 3b), peaks at 685, 765, 1075, and 1140 cm^{-1} disappear, and we can conclude that character of these spectral peaks correspond to modes with large displacements along the x-axis (e.g., OO stretch, OH wagging, OH rocking, or BH stretch). We will later confirm these assignments using the position−position correlation function and INM analysis (see below).

Normal mode analysis predicts two BH bend modes at 1433 and 1686 cm^{-1} at the MP2 level. Quantum studies[15] place these modes between 1300 and 1500 cm^{-1}. Our MP2 spectrum (Figure 2b) does not show distinct spectral peaks in this region. However, when MP2 IR intensities are calculated using the perpendicular component of the dipole−dipole correlation funcion, BH bend modes can be found in the spectrum at 1450 and 1650 cm^{-1} (Figure 3b). Both spectral peaks shift by about 450 cm^{-1} upon isotopic substitution.

B. Position−Position Spectra. In order to address the issue with the many overlapping features in the IR spectrum shown in the Figure 2b, we calculated a few representative position−position spectra from the coordinates saved along MP2 trajectory at 100 K. Figure 4 shows the total intensities and the parallel and perpendicular components along the x- and yz-directions, respectively, for the terminal hydrogens, bridging hydrogen, and for OO-stretch position vectors. The placement of the peaks in the position−position spectra can be directly compared to the placement of peaks in the IR

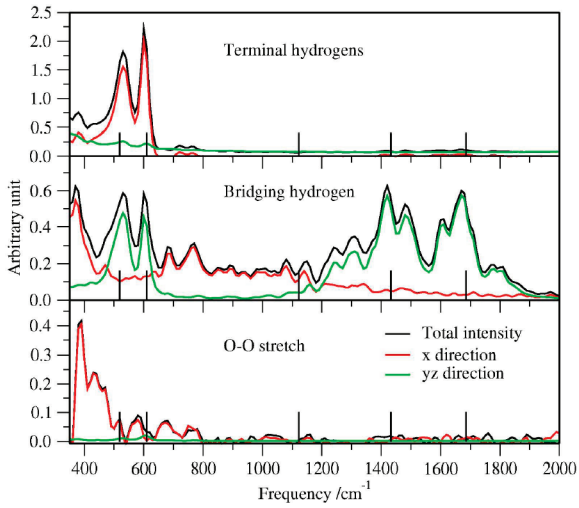


Figure 4. The MP2 vibrational spectrum at 100 K obtained by Fourier transform of the position−position correlation function. The spectrum was decomposed into the contibutions from the terminal hydrogens H3 and H5 (top panel), bridging hydrogen H1 (middle panel), and OO stretch position vectors (bottom panel). This figure shows the total intensities (black lines) and the parallel (red lines) and perpendicular components (green lines) along the *x*- and *yz*-directions, respectively.

spectra (Figure 2b). However, the intensity of the peaks should not be compared directly, because a position−position spectrum does not reflect dipole activity.

The position−position spectra clearly identify the activity of OH wag, OH rock, OO stretch, BH parallel, and perpendicular vibrations. We note that the base lines of these spectra are well-defined. Our MD simulations predict the OO stretch vibration to be blue-shifted by about 90 cm^{-1}. The position−position spectrum is showing the OO stretch peak at 400 cm^{-1}, while the harmonic frequency is 310 cm^{-1}. The terminal hydrogens exhibit activity in the spectral region from 400 to 2000 cm^{-1} for OH wag and OH rock vibrations, as expected from the normal-mode analysis. The position−position spectrum of the bridging hydrogen is more complex, clearly showing regions where the parallel (along *x*-axis) and perpendicular (along *y*- and *z*-axes) components dominate. BH perpendicular component is showing two doublets near two bending harmonic vibrations at 1433 and 1686 cm^{-1} and two weak features at 1250 and 1300 cm^{-1}. The BH perpendicular component is also strongly coupled to both OH wag and OH rock vibrations at 519 and 610 cm^{-1}, respectively.

The BH mode exhibits large displacements along the O−O axis (*x*-axis) up to 1200 cm^{-1}. The harmonic frequency for BH stretch along the *x* axis is 1120 cm^{-1} (Table 1). The baseline of the position−position spectrum of parallel components is not as well-defined as that of the perpendicular component. There is a strong peak at 400 cm^{-1} that coincides with the O−O stretch and four weak features at 700, 780, 1080, and 1140 cm^{-1} that are most likely combination bands of BH stretch and other modes. This analysis of the position−position spectra brings new insights into understanding the complexity of the IR spectrum shown in Figure 2b).

Comparison of the MP2 and experimental spectra demonstrates that MP2 is capturing the essential spectral feature of the BH stretch mode between 600 and 900 cm^{-1}. The main difference between our MP2 spectrum and experimental spectra[16,21] lies in the observation of the feature near 1090 cm^{-1}. Quantum studies[15,16] reveal large displacements of the shared proton along and/or perpendicular to the O−O axis in the 1000−1100 cm^{-1} region. From the analysis of our IR spectrum using the polarized dipole absorptions and position−position correlation function, we can conclude that these spectral features most likely originate from the parallel motion of the shared proton coupled to other modes (Figure 2). Our results predict the perpendicular motion of the shared proton between 1250 and 1650 cm^{-1} (Figure 4).

C. Comparison of Methods. We also examined how B3LYP and B3PW91 methods perform for this system. The DFT and MP2 structural parameters, potential energy features, the results from the direct MD simulations, and interpretation of the trajectories were compared with the aim of assessing the reliability of DFT functionals for this anion. The B3LYP and B3PW91 structural parameters are listed in the Table 1. The geometry optimized at the B3LYP level of theory shows a nonplanar structure with an asymmetric hydrogen bond, while the B3PW91 calculations failed to reproduce the asymmetry of the hydrogen bond in $H_3O_2^-$. If the larger basis set 6-311++G(2d,2p) is used, $H_3O_2^-$ remains nearly symmetrical with R_1 and R_2 distances equal to 1.227 and 1.214 Å, respectively. Harmonic vibrational frequencies for $H_3O_2^-$ and $D_3O_2^-$ are listed in the Table 2. B3LYP and MP2 vibrational frequencies are in very good agreement. The largest differences appear in the BH modes. BH stretch frequencies along the *x* axis are 967 and 1120 cm^{-1} at B3LYP and MP2 levels, respectively. The bridging hydrogen motion is a large amplitude mode with normal mode displacement values above 0.8. The BH stretch frequency is sensitive to the shape of the potential near the transition state for the proton transfer. The normal mode assignment for B3PW91 differs from the B3LYP and MP2 mode assignment because B3PW91 failed to reproduce the asymmetry of the hydrogen bond in $H_3O_2^-$. In the B3LYP and MP2 case, the OH stretch frequencies are localized on either of the OH bonds, while in the B3PW91 case, OH stretch frequencies are assigned as symmetric and asymmetric OH stretch.

The proton motion between the two oxygens in $H_3O_2^-$ is characterized by a 74 cm^{-1} barrier on the potential energy surface calculated at CCSD(T) computational level.[14] MP2 and B3LYP calculations predict much lower barrier heights about 13 and 5 cm^{-1}, respectively, and $H_3O_2^-$ experiences no energy barrier at the B3PW91 level. The qualitative differences between the potentials calculated at DFT and MP2 levels are further analyzed in terms of the potential energy surface scanned along the normal mode displacements. Figure 5a shows potential energy curves along the BH stretch and O−H stretch normal mode displacements. The shapes of the potential energy curves are almost identical for B3LYP and MP2 calculations, while at the B3PW91 level the potential is much more flat at displacements beyond 0.6 Å. Figure 5b depicts the

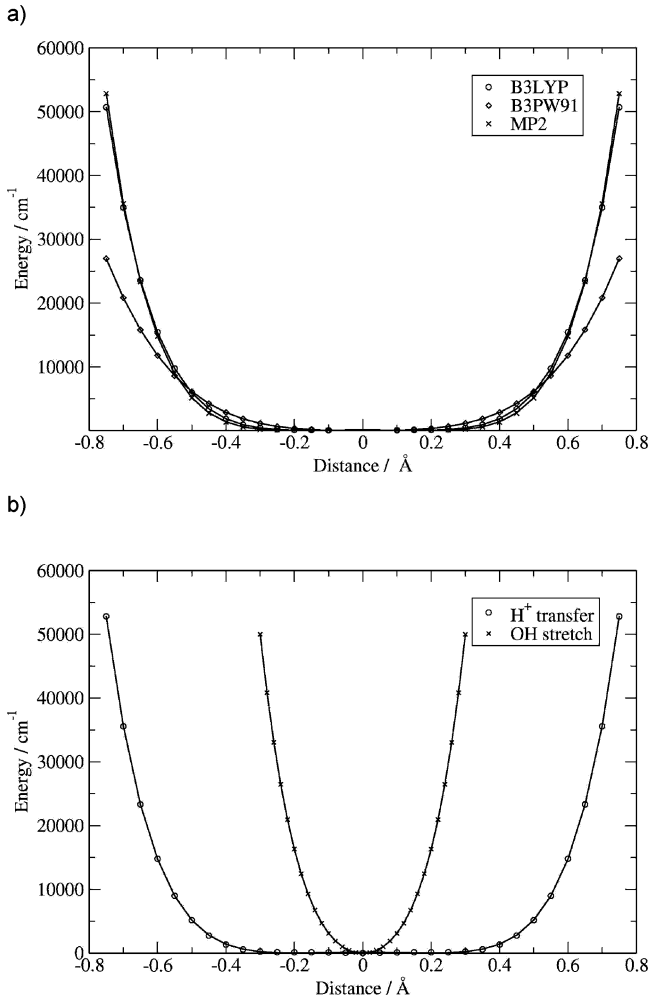


Figure 5. Potential energy curves along (a) the H^+ transfer (BH stretch along *x*-axis) normal mode displacements calculated by B3LYP, B3PW91, and MP2 methods. (b) H^+ transfer and O−H stretch normal mode displacements calculated by the MP2 method.

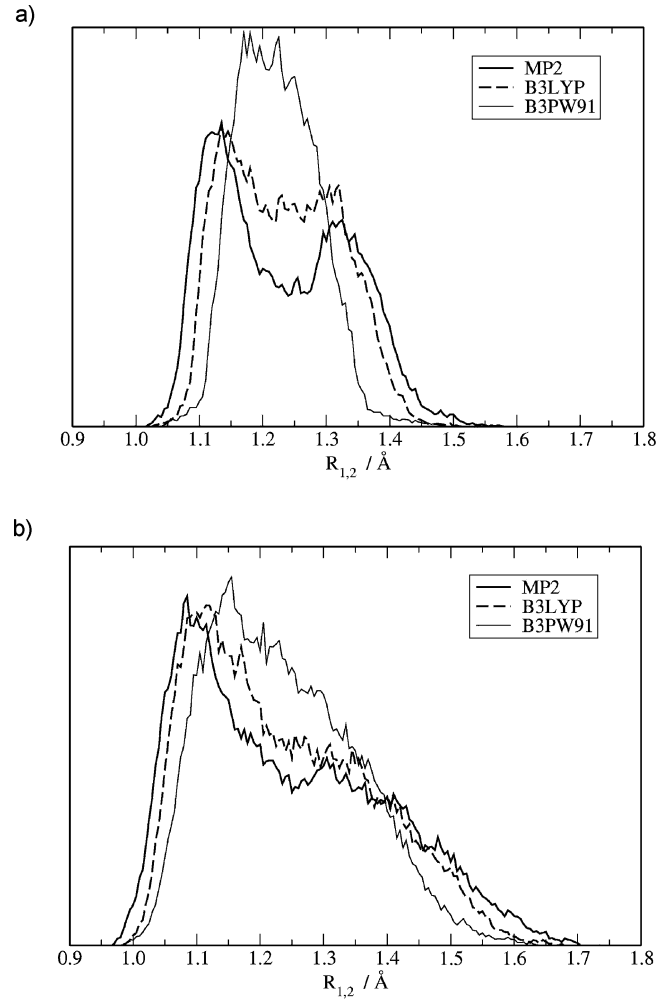


Figure 6. Bond-length distributions for H^+ transfer according to MD B3LYP, B3PW91, and MP2 calculations at (a) 100 K and (b) 300 K.

floppiness of molecular motion along the BH stretch normal mode displacement compared to the OH stretch mode at the MP2 level of theory.

The bond-length distributions for a proton transfer averaged over five direct MD trajectories calculated by B3LYP, B3PW91, and MP2 methods are shown in Figure 6. It is apparent that difference in the potential energy surface near the transition state for the proton transfer in $H_3O_2^-$ influences the dynamical behavior of the system. At B3LYP and MP2 levels the bond-length distributions are broad and show two peaks, while at the B3PW91 level it is much narrower due to the absence of the barrier for the proton transfer.

Figure 7 presents a comparison of the simulated IR spectra up to 2000 cm^{-1} calculated at 100 K by B3LYP, B3PW91, and MP2 methods. The spectral intensities from MP2 and B3LYP calculations are very similar, while B3PW91 predicts different spectral intensities due to the qualitative differences in the potential energy and dipole moment surfaces.

D. Temperature Dependence of the B3LYP Spectra. Since it is not entirely clear from the experiment[21] at which temperature the spectrum was measured, we performed a

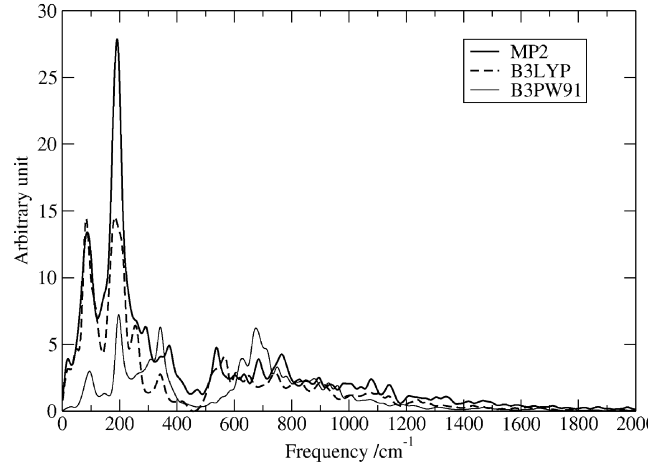


Figure 7. Comparison of MD B3LYP, B3PW91, and MP2 IR spectra for $H_3O_2^-$ at 100 K.

series of calculations to try to determine experimental conditions. In Figure 8 we show the temperature dependence of the B3LYP spectra from 10 to 300 K. At 10 K the B3LYP spectrum is essentially harmonic. Only the BH stretch vibration along the *x*-axis is red-shifted by about 50 cm^{-1}. At 50 K the spectrum is starting to show a significant

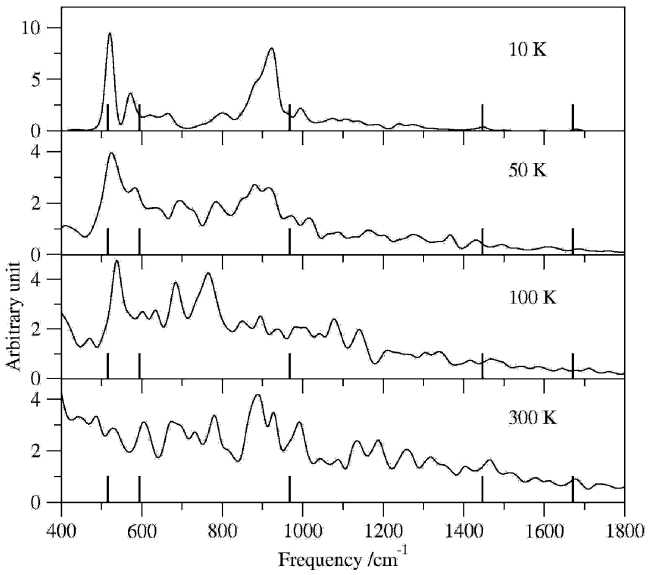


Figure 8. Temperature dependence of the B3LYP spectra from 10 to 300 K. The IR spectra were obtained from 10 trajectories propagated for 10 ps. The harmonic frequencies are shown as sticks in the spectrum.

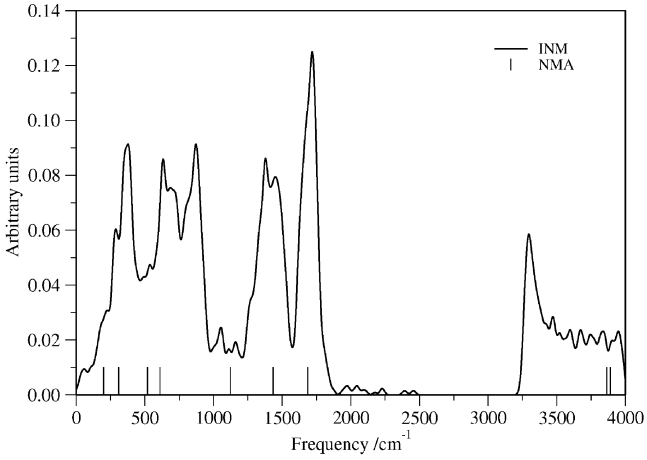


Figure 9. Vibrational density of states obtained from INM analysis of MP2 MD simulations at 100 K. The harmonic frequencies are shown as sticks in the spectrum.

temperature effect. One can see new peaks arising between 600 and 900 cm^{-1} that originate from coupling of the BH with other modes, as we have mentioned above. This trend continues at higher temperature, where there is also much more activity above 1000 cm^{-1}. On the basis of these observations we believe that the experimental spectrum[21] was measured at a relatively low temperature, possibly <50 K.

E. INM Spectra. The INM spectrum was obtained by calculating the INM frequencies for 500 instantaneous configurations drawn from the course of a direct MD simulations at the MP2 level at temperature 100 K. Because the system is not at its global potential energy minimum, some modes represent unstable motions, and these unstable modes have imaginary frequencies, which are omitted in the analysis. Figure 9 shows the INM density of states in the range from 0 to 4000 cm^{-1} for $H_3O_2^-$. The normal-mode frequencies are shown as sticks in the spectrum. In this figure we can see broad bands peaking at 375, 630, 870, 1400, and

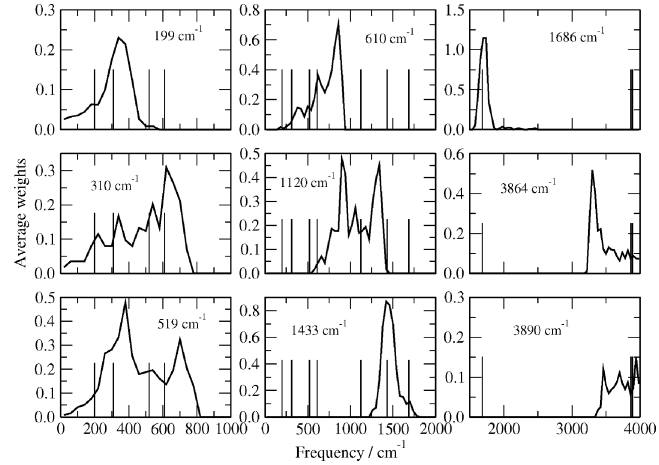


Figure 10. Average weights representing the overlap of the MP2 instantaneous normal mode vectors and the normal mode vectors calculated at the $H_3O_2^-$ minimum structure. The harmonic frequencies are shown as sticks in the spectrum.

1720 cm^{-1}. The peak positions of the density of states can be identified as normal-mode frequencies at 310, 610, 1433, and 1686 cm^{-1}. However, there is no normal mode in the spectral region between 610 and 1120 cm^{-1}; therefore, we assume that the peak at 870 cm^{-1} is either a combination band or a strongly shifted fundamental. Instantaneous normal modes for OH stretch motion are red-shifted with the maximum peak at 3300 cm^{-1}. MP2 MD simulations predict two weak OH stretch frequencies at 3685 and 3880 cm^{-1} for $H_3O_2^-$ and 2660 and 2820 cm^{-1} for $D_3O_2^-$. However, the position of the peaks are not thought to be reliable because of the relatively long propagation time step.

Figure 10 shows average weights for all nine $H_3O_2^-$ normal modes. The peaks in the spectral regions around 1400 and 1700 cm^{-1} are relatively narrow. We conclude that the instantaneous modes in these spectral windows overlap with normal modes at 1433 and 1686 cm^{-1}, respectively. Instantaneous normal modes up to 800 cm^{-1} overlap with several low frequency normal modes, showing blue and red anharmonic shifts. BH stretch normal mode at 1120 cm^{-1} overlaps with instantaneous normal modes in the range from 700 to 1400 cm^{-1} showing maximum overlaps at 900 and 1340 cm^{-1}. Experimentally, BH stretch was observed at 697 cm^{-1}.[21] The character of the broad spectral feature in the INM density of states (Figure 9) near 870 cm^{-1} can be assigned to the combination of the OH rock vibration (75%) and BH stretch vibration along the *x*-axis (25%). The assignment of this frequency is in agreement with the assignment of the IR spectrum obtained from our direct MD simulations.

INM analysis that we carried out in this work is temperature-dependent, because the instantaneous structures were sampled from MD trajectories propagated at fixed temperature. At very low temperature, MD trajectories sample structures that are very similar to the minimum on the potential energy surface, while at higher temperature, multiple minima can be visited; therefore, the instantaneous normal mode spectrum could be regarded as a weighted average over several minima. In such a case, overlaps of normal mode vectors for instantaneous struc-

tures with other minima can be weighted by a factor that reflects the distance of an instantaneous structure from that minimum, i.e., $1/|\vec{r}^{MIN} - \vec{r}^{INM}|^2$. We plan to use this approach for studying larger, more floppy water clusters.

4. Conclusions

In this work, we present IR spectra of $H_3O_2^-$ and its isotopomer $D_3O_2^-$ using the MP2, B3LYP, and B3PW91 direct molecular dynamics simulations. Despite some qualitative differences in the interaction potentials between B3PW91 and MP2 levels, the spectral patterns were found to be in reasonable agreement. IR spectra generated by B3LYP and MP2 methods show a qualitatively similar picture as those obtained from experimental studies.[16,21] We assigned the spectral peaks in our classical IR spectrum calculated at the MP2 computational level at 100 K using INM analysis, by decomposition the IR spectral intensities into the parallel and perpendicular components, and by analysis of the position−position correlation function. Our results predict the bridging hydrogen stretch between 600 and 900 cm^{-1} and bridging hydrogen bend vibrations between 1250 and 1650 cm^{-1}. On the basis of our temperature-dependence study of the B3LYP spectrum, we believe that the experimental spectrum[21] was measured at a very low temperature.

With this study, we hope that we have presented convincing evidence that the B3LYP method provides an accurate description of the interaction potential and dipole surfaces and can be used in simulations of larger hydrated ions. For more concrete assignment of spectral features, we plan to carry out driven MD simulations.[23,43] Spectra will be analyzed using the method shown in ref 23.

Acknowledgment. We thank the Research Corp. for support of this work, Kennesaw State University for support under Faculty Summer Research Award, Incentive Funding Awards. We also thank Prof. J. M. Bowman (Emory University, Atlanta, GA) and Prof. M. B. Mitchell (Kennesaw State University) for many useful discussions on this project.

References

(1) Nagle, J. F. *J. Bioenerg. Biomembr.* **1987**, *19*, 413–426.

(2) Xantheas, S. S. *J. Am. Chem. Soc.* **1995**, *117*, 10373–10380.

(3) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *Science* **1997**, *275*, 817–820.

(4) Tunon, I.; Martins-Costa, M. T. C.; Millot, C.; Ruiz-Lopez, M. F. *J. Chem. Phys.* **1997**, *106*, 3633–3642.

(5) Perez del Valle, C.; Novoa, J. J. *Chem. Phys. Lett.* **1997**, *269*, 401–407.

(6) Ruiz-Lopez, M. F.; Oliva, A.; Bertran, J. *J. Phys. Chem.* **1998**, *102*, 10728–10735.

(7) Turki, N.; Milet, A.; Rahmouni, A.; Ouamerali, O.; Moszynski, R.; Kochanski, E.; Wormer, P. E. S. *J. Chem. Phys.* **1998**, *109*, 7157–7168.

(8) Wei, D.; Proynov, E. I.; Milet, A.; Salahub, D. R. *J. Phys. Chem. A* **2000**, *104*, 2384–2395.

(9) Barich, D. H.; Nicholas, J. B.; Haw, J. F. *J. Phys. Chem. A* **2001**, *105*, 4708–4715.

(10) Chaudhuri, C.; Wang, Y. S.; Jiang, J. C.; Lee, Y. T.; Chang, H. C.; Nieder-Schatteburg, G. *Mol. Phys.* **2001**, *99*, 1161–1173.

(11) Samson, C. C. M.; Klopper, W. *J. Mol. Struct. (THEOCHEM)* **2002**, *586*, 201–208.

(12) Tuckerman, M. E.; Marx, D.; Parrinello, M. *Nature* **2002**, *417*, 926–929.

(13) Yeh, L. I.; Okumura, M.; Meyers, J. D.; Price, J. M.; Lee, Y. T. *J. Chem. Phys.* **1989**, *91*, 7319–7330.

(14) Huang, X.; Braams, B. J.; Carter, S.; Bowman, J. M. *J. Am. Chem. Soc.* **2004**, *126*, 5042–5043.

(15) McCoy, A. B.; Huang, X.; Carter, S.; Bowman, J. M. *J. Chem. Phys.* **2005**, *123*, 064317−064330.

(16) Diken, E. G.; Headrick, J. M.; Roscioli, J. R.; Bopp, J. C.; Johnson, M. A.; McCoy, A. B.; Huang, X.; Carter, S.; Bowman, J. M. *J. Phys. Chem. A* **2005**, *109*, 571–575.

(17) Teige, V. E.; Li, X.; Iyengar, S. S. *J. Phys. Chem. A* **2007**, *111*, 4815–4820.

(18) Yang, Y.; Kuhn, O. *Z. Phys. Chem.* **2008**, *222*, 1375–1387.

(19) Price, E. A.; Robertson, W. H.; Diken, E. G.; Weddle, G. H.; Johnson, M. A. *Chem. Phys. Lett.* **2002**, *366*, 412–416.

(20) Robertson, W. H.; Diken, E. G.; Price, E. A.; Shin, J. W.; Johnson, M. A. *Science* **2003**, *299*, 1367–1372.

(21) Diken, E. G.; Headrick, J. M.; Roscioli, J. R.; Bopp, J. C.; Johnson, M. A.; McCoy, A. B. *J. Phys. Chem. A* **2005**, *109*, 1487–1490.

(22) Hammer, N. I.; Diken, E. G.; Headrick, J. M.; Roscioli, J. R.; Johnson, M. A.; Myshakin, E. M.; Jordan, K. D.; McCoy, A. B.; Huang, X.; Bowman, J. M.; Carter, S. *J. Chem. Phys.* **2005**, *122*, 244301–244310.

(23) Kaledin, M.; Kaledin, A. L.; Bowman, J. M. *J. Phys. Chem. A* **2006**, *110*, 2933–2939.

(24) Kubelka, J.; Keiderling, T. A. *J. Phys. Chem. A* **2001**, *105*, 10922–10928.

(25) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw-Hill: New York, 1955.

(26) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503–506.

(27) Saebo, M. J. S.; Almlöf, J. *Chem. Phys. Lett.* **1989**, *154*, 83–89.

(28) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 275–280.

(29) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(30) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(32) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

(33) 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 C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(34) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(35) Haile, J. M. *Molecular Dynamics Simulation: Elementary Methods*; Wiley & Sons: New York, 1992; p 277.

(36) Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1981**, *74*, 4872–4882.

(37) Stratt, R. M. *Acc. Chem. Res.* **1995**, *28*, 201–207.

(38) Keyes, T. *J. Chem. Phys.* **1996**, *104*, 9349–9356.

(39) Buchner, M.; Ladanyi, B. M.; Stratt, R. M. *J. Chem. Phys.* **1992**, *97*, 8522–8535.

(40) Cho, M.; Fleming, G. R.; Saito, S.; Ohmine, I.; Stratt, R. M. *J. Chem. Phys.* **1994**, *100*, 6672–6683.

(41) Li, X.; Moore, D. T.; Iyengar, S. S. *J. Chem. Phys.* **2008**, *128*, 184308.

(42) Sauer, J.; Dobler, J. *ChemPhysChem* **2005**, *6*, 1706–1710.

(43) Bowman, J. M.; Zhang, X.; Brown, A. *J. Chem. Phys.* **2003**, *119*, 646–650.

CT8004485

JCTC Journal of Chemical Theory and Computation

Influence of Adsorbate-Free Atoms on Δ-XANES Signatures

Stanislav Stoupin*

Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115

Received December 8, 2008

Abstract: The use of differential X-ray Absorption Near Edge Spectroscopy (Δ-XANES) for analysis of site specific adsorption on metallic electrodes relies on theoretical Δ-XANES signatures for analysis of experimental Δ-XANES fingerprints. A simple model, currently used in the analysis, considers changes in X-ray absorption properties of adsorbing atoms only. This model has been extended to include changes in X-ray absorption for other atoms of the same type that remain adsorbate-free. Configurational averaging has been applied to calculate difference spectra of a Pt_6 cluster with an oxygen atom adsorbed at different sites. The extended theory shows that contribution of the adsorbate-free atoms might become significant as it affects the shape profiles of the theoretical signatures. The effect, most prominent at the absorption edge energy, is interpreted in terms of change in the electronic structure of the cluster due to oxygen adsorption. In addition to model dependence of the theoretical signatures, challenges to the application of Δ-XANES to the experimentally obtained fingerprints are discussed.

Introduction

Knowledge of adsorbate binding sites is essential for understanding catalytic and electrochemical reactions. For instance, oxygen adsorption on a complex surface of a high dispersion Pt based catalyst is an elementary step in both anode and cathode fuel cell reactions. Oxygen participation in these reactions depends on a large number of parameters including electrode potentials, adsorbate content, catalyst composition, and morphology. This situation alone generates high demand for reliable methods of surface adsorbate analysis under reaction conditions. Numerous studies report systems and conditions where a particular type of catalyst-adsorbate interaction prevails. However, only a few in situ methods provide structural and/or electronic information from which details of adsorbate-catalyst interaction can be inferred. X-ray Absorption Spectroscopy (XAS) is one of the most valuable techniques that serve the purpose.[1] XAS at the Pt L_3 X-ray absorption edge is widely used for in situ characterization of Pt-based catalysts.[2−14]

An emerging strategy for the analysis of site-specific adsorption is differential X-ray Absorption Near Edge Spectroscopy (Δ-XANES).[8−14] XANES of a catalyst under adsorbate-free conditions is subtracted from XANES under conditions conducive to adsorption (e.g., by altering the potential or chemical environment). The Δ-XANES is sensitive to changes in metal-adsorbate and metal−metal interactions. The interpretation of the experimental Δ-XANES (Δμ-fingerprint) is at an early stage of development. Ramaker et al. calculate theoretical XANES of model clusters (e.g., Pt_6 Janin cluster[15]) with and without adsorbates using a full-multiple scattering code (FEFF8), which employs self-consistent field calculations of the local electronic structure.[16,17] The corresponding theoretical Δ-XANES, hereafter referred to as Δμ-signatures, are used for analysis of Δμ-fingerprints.[8−14] The Δμ-fingerprints are correlated to adsorbate coverage profiles by plotting magnitudes of certain characteristic peaks versus parameters of interest (e.g., electrode potential, adsorbate concentration).

To date, the Δμ-signatures for oxygen adsorption in the four oxygen-adsorption configurations only include X-ray absorption events from a single adsorbing Pt atom (atom-0 in Figure 1) of the Janin cluster at the Pt L_3 edge.[9] Further, these will be referred to as *limited-adsorber* Δμ-signatures.

* Corresponding author e-mail: sstoupin@aps.anl.gov. Current address: Advanced Photon Source, Argonne National Laboratory, IL 60439.

10.1021/ct800544a CCC: $40.75
Published on Web 04/02/2009

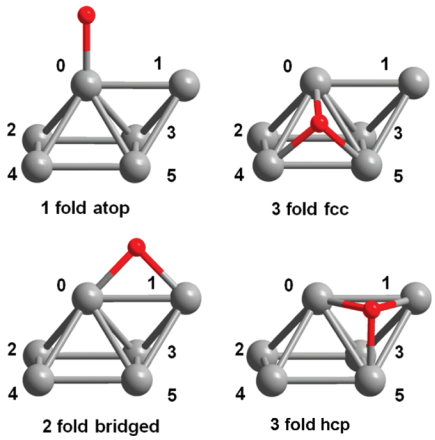


Figure 1. Oxygen adsorption configurations of the Janin Pt_6 cluster and numbering of the Pt atoms.

Analysis of experimental fingerprints using the *limited-adsorber* signatures is based on an assumption that the presence of adsorbate does not alter X-ray absorption properties of the adsorbate-free atoms of an electrochemical system under consideration. This assumption seems to be applicable in the case of low adsorbate coverage where a small number of adsorbed species does not significantly disturb the electronic structure of the system. However, it is not obvious in a case where the number of adsorbed species is comparable to the number of atoms in the catalytic nanoparticle.

For example, Ankudinov et al.[18] showed that changes in the near-neighbor Pt scattering potentials due to adsorbate-Pt interaction (i.e., changes in the Pt electronic structure) may dominate over the effect of Pt-adsorbate photoelectron multiple scattering. Furthermore, Bazin et al. discuss the importance of configurational averaging when calculating XANES of small clusters.[19] It was illustrated that per-atom Pt L_3 edge XANES differ among nonequivalent Pt sites. In this regard, presence of a single adsorbate on the six-atom Pt Janin cluster can be considered as a case of intermediate coverage.

Because XAS probes all cluster atoms, a more rigorous theoretical treatment of Δ-XANES requires inclusion of X-ray absorption events originating from all cluster atoms (i.e., *all-atoms* model). In this work, electronic properties of the Janin cluster in the four oxygen-adsorption configurations are correlated to the shape profiles of atom specific $\Delta\mu$-signatures based on results of FEFF8 calculations. It is shown that the shape profiles of *all-atoms* $\Delta\mu$-signatures differ from the *limited-adsorber* $\Delta\mu$-signatures currently used for Δ-XANES analysis of oxygen adsorption. Limitations on the applicability of both models are discussed.

Computational Approach

The Pt L_3 XANES white-line is a strong peak due to unoccupied Pt d-densities of states (d-DOS) near the photoelectron threshold energy (i.e., Fermi level). The FEFF8 code default Hedin-Lundqvist self-energy (omitted EXCHANGE card) adequately reproduces the Pt white-line.[20−23] The final states calculated without the presence of a screened core-hole (NOHOLE card) show better agreement with experimental data.[22−24] An atomic potential (IPOT) was assigned to the adsorbed oxygen and every Pt atom of the Janin cluster. Thus, the number of unique atomic potentials in each calculation was 6 for the adsorbate-free (clean) cluster and 7 for the clusters with adsorbed oxygen. It is noted that the number of unique potentials can be reduced for the clean, 1-fold atop and 2-fold bridged configurations due to their symmetry.

The version of the code used in this work was FEFF8.4. The Pt L_3 XANES spectrum was calculated for each Pt atom by assignment of the X-ray absorber atomic potential (IPOT 0) to that atom. This procedure was applied to all configurations of the Janin cluster using the four oxygen adsorption sites as previously reported (Figure 1). The signatures were computed on a per-atom basis using Athena program.[25] The absorption edge energy (11564 eV) was assigned to the inflection point of the calculated XANES for atom-0 of the adsorbate-free cluster, which resulted in 2.6 eV shift of the FEFF8 absolute energy scale. Thus, the code underestimates the absolute edge energy if defined as the inflection point. The shift was applied to all calculated spectra.

Results

The white-line intensities for atom-0 increase in any of the four oxygen-adsorption configurations, which is accompanied by reduction in the population of the atom-0 d-states (Figure 2a). The subtraction of the atom-0 clean cluster spectrum from the atom-0 atop or n-fold adsorption configurations yields the *limited-adsorber* signatures with no derivative-like appearance (i.e., signatures of Ramaker et al.[9−11] reproduced in our calculations). In contrast, only a minor change in the population of the d-states occurs for the adsorbate-free atoms. For these atoms, the variation in the white-line intensity with oxygen-adsorption configuration is small relative to that of any of the adsorbing atoms. Figure 2b shows XANES of the clean cluster atom-2 and that of the atom-2 in the 3-fold fcc configuration. The shift to higher energy of the 3-fold XANES results in the derivative-like appearance of the atom-2 $\Delta\mu$-signature (inset in Figure 2b).

The shape of the XANES relates to FEFF8 calculated Pt local d-DOS (d-LDOS) spectra above the Fermi level. Figure 2c contrasts d-LDOS of atom-0 in the clean cluster (blue line) with that of the atom-0 in the atop configuration (red line). Although the number of vacant states of the atop atom-0 is higher just above the Fermi level, the density of states reduces at higher energies and becomes lower than that of atom-0 in the clean cluster. This reduction is correlated with intensity of XANES of the atom-0 (Figure 2a). Similarly, the variation of d-LDOS of atom-2 in the 3-fold fcc configuration (Figure 2d − purple line) vs d-LDOS of atom-2 of the clean cluster (blue line) is correlated with the shapes of atom-2 XANES (Figure 2b). In general, for every Pt atom in the clusters, self-consistent field calculations of the local electronic structure yield variation of d-LDOS spectra upon oxygen adsorption. These variations correlate with the shape and intensity of the calculated atom-specific XANES/Δ-XANES and are rather significant for adsorbate-free atoms of the Janin cluster.

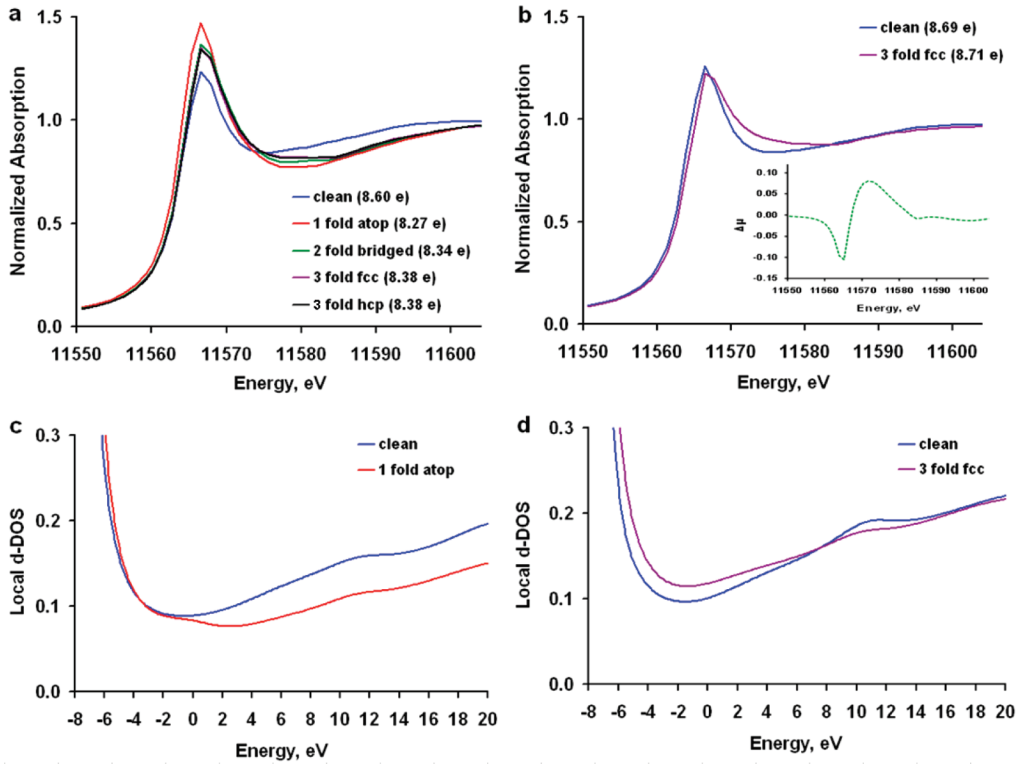


Figure 2. Variation of the calculated atom specific Pt L_3 edge XANES and d-LDOS with oxygen adsorption configuration: (a) XANES of adsorbing atom-0: the dominant variation is the increase in the white-line intensity with reduction in the population of the d states (the populations are shown in the legend). (b) XANES of adsorbate-free atom-2 in the 3-fold fcc configuration (purple) and of atom-2 in the clean cluster (blue). The inset shows the result of subtractive normalization of the purple line to the blue line (i.e., atom-2 $\Delta\mu$-signature in the 3-fold fcc configuration). (c) d-LDOS of atom-0 in the atop configuration (red) and of atom-0 in the clean cluster (blue). (d) d-LDOS of adsorbate-free atom-2 in the 3-fold fcc configuration (purple) and of atom-2 in the clean cluster (blue).

Figure 3a-d summarizes the $\Delta\mu$-signatures on a per-atom basis in all configurations. The blue lines are the average of the per-atom signatures (*all-atoms* $\Delta\mu$-signatures) for each of the adsorption configurations. Averaging is for illustrative purposes to mimic the Δ-XANES fingerprint of the Janin cluster as if it was a real adsorption system (i.e., all Pt atoms are equally probed by the X-rays). With the exception of the 1-fold atop configuration, the *all-atoms* $\Delta\mu$-signatures have a derivative-like appearance. The red lines (Figure 3a-d) are the *limited-absorber* $\Delta\mu$-signatures based on X-ray absorption events originating only from one adsorbing atom (atom-0). The shape profiles and intensities of the *all-atoms* $\Delta\mu$-signatures (blue lines) are compared to the respective *limited-adsorber* signatures (red lines) (Figure 3a-d). FEFF8 calculated charge transfer magnitudes, d-LDOS populations (on a per-atom basis), and Fermi levels of the clean cluster, the 1-fold atop, and the n-fold cluster configurations are summarized in Table 1.

Discussion

The factors that determine the shape and intensity of the atom-specific signatures at low energy side (i.e., in vicinity of the absorption edge) are variations in the white-line intensity and changes in the absorption edge energy. As noted earlier, the primary factor for adsorbing atoms is the increase in the white-line intensity correlated to the reduction in the population of the d-LDOS due to charge transfer effects (e.g., atom-0 - Figure 2a). In the following discussion, the influence of changes in the absorption edge energy is analyzed.

Fermi level of any given cluster sets the photoelectron threshold for all atoms in that cluster (Table 1). With the exception of the 1-fold atop configuration, the Fermi level shifts positively for all other clusters with an adsorbed oxygen atom (n-fold configurations). Thus, a positive shift of the absorption edge, leading to a derivative-like appearance, is expected for every Pt atom in the n-fold configurations. Although the direction of this shift is correlated with the shift of the Fermi level, it is noted that the absolute energy of the absorption edge is an estimate. The code estimates absolute energies by calculating the difference in the total relativistic Dirac/Fock energy of an atom with and without a core-hole. Under an assumption that the absolute energies are equally underestimated for all atoms, the shift of the absorption edge energy is due to the shift in the Fermi level.

The loss of intensity at the edge region due to the positive shift of the absorption edge is compensated by an increase in the white-line intensity upon adsorption of oxygen. This compensation can be complete (e.g., $\Delta\mu$ of atom-0 in all n-fold configurations − Figure 3a-d) or incomplete, such that a small negative peak is still present (e.g., $\Delta\mu$ of atom-4 and atom-5 in Figure 3b). In contrast, the negative peak at the low energy side is substantial in the adsorbate-free atom signatures because the loss of intensity due to the absorption edge shift is not countered by an increase in the white-line

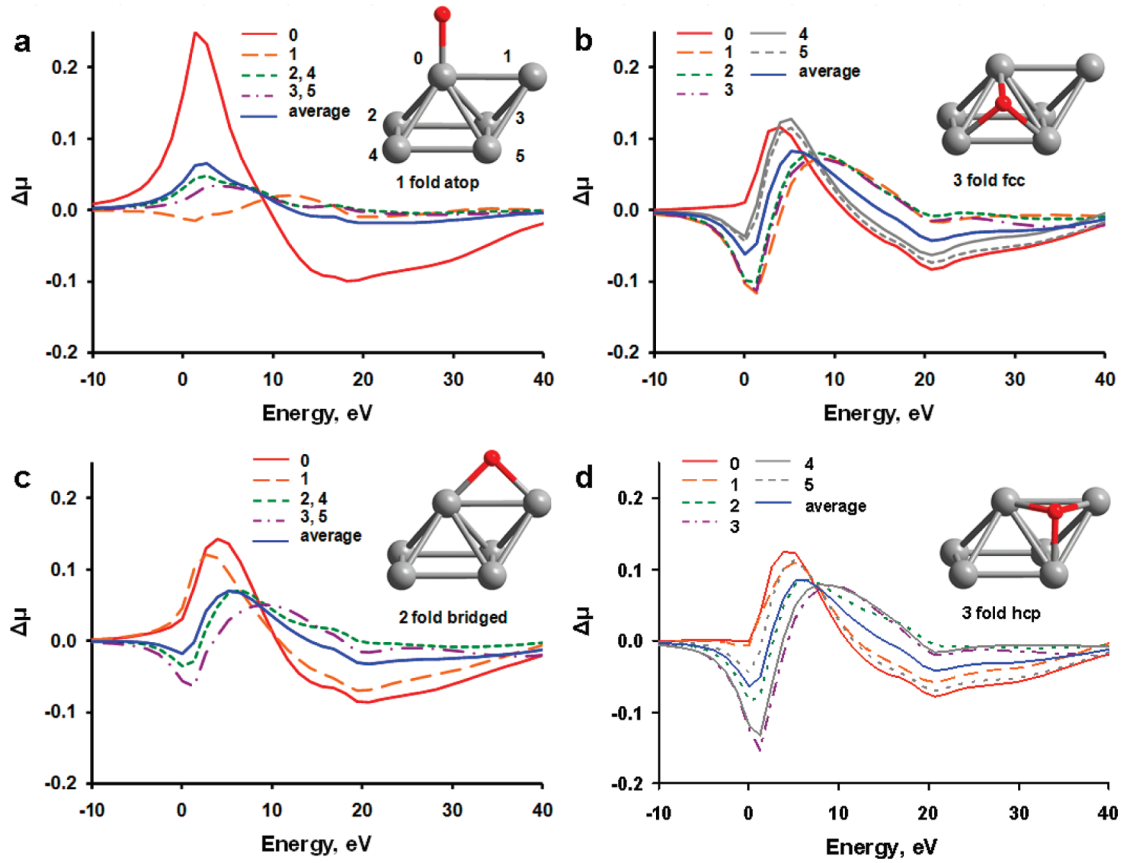


Figure 3. FEFF8 $\Delta\mu$ signatures on a per-atom basis calculated for Pt_6 Janin cluster with O adsorbed on the (a) 1-fold atop, (b) 3-fold fcc, (c) 2-fold bridged, and (d) 3-fold hcp sites. The red lines are signatures of Ramaker et al. reproduced in our calculations (*limited-adsorber* signatures). The blue lines are the average value $\Delta\mu$-signatures (*all-atoms* signatures).

Table 1. FEFF8 Atomic Partial Charge in Units of the Elementary Charge, Atomic d Electron Counts, and Cluster Fermi Levels

atom/cluster	clean	1-fold atop	2-fold bridged	3-fold fcc	3-fold hcp
O	–	−0.246	−0.095	0.012	0.006
Pt0	0.051(8.60)	0.320(8.27)[a]	0.262(8.34)[a]	0.239(8.38)[a]	0.223(8.38)[a]
Pt1	−0.030(8.70)	−0.022(8.70)	0.239(8.41)[a]	−0.232(8.72)	0.167(8.43)[a]
Pt2	−0.033(8.69)	−0.044(8.67)	−0.133(8.69)	−0.240(8.71)	−0.225(8.70)
Pt3	0.023(8.62)	0.018(8.62)	−0.070(8.64)	−0.158(8.65)	−0.154(8.68)
Pt4	−0.033(8.69)	−0.044(8.67)	−0.133(8.69)	0.171(8.44)[a]	−0.223(8.73)
Pt5	0.023(8.62)	0.018(8.62)	−0.070(8.64)	0.208(8.42)[a]	0.207(8.42)[a]
E_F	−7.135	−7.198	−7.021	−6.830	−6.816

[a] Adsorbing Pt atoms.

intensity. Thus, the derivative-like appearance of the n-fold *all-atoms* signatures of the Janin cluster is characteristic to its adsorbate-free atoms.

An important remark is that the offered interpretation of the derivative-like shape is not applicable to the case of hydrogen adsorption on Pt. The effect of H adsorption on Pt L_3 edge XANES has been extensively studied.[8,18,26] A substantial negative contribution was observed experimentally and modeled by *limited-adsorber* signatures for H in the n-fold configurations. In order to validate the *all-atoms* model, the per-atom $\Delta\mu$-signatures were calculated for H atom adsorbed on the Janin cluster in the four configurations. The results are presented in the Supporting Information. In contrast to the case of O, the shifts of the Fermi level and the charge transfer effects are insufficient to explain the negative contribution. The resulting $\Delta\mu$-signatures of the adsorbate-free atoms of the n-fold configurations are similar to the n-fold *limited-adsorbate* signature reported earlier. Thus, in the case of hydrogen, configurational averaging does not produce new features in the spectral shape of the theoretical Δ-XANES.

If considered as theoretical standards, the calculated $\Delta\mu$-signatures of the adsorbate-free atoms are likely exaggerated due to the small size of the cluster. For example, realistic electrochemical Pt/C clusters are usually about 1.5−2 nm in size (100−300 atoms assuming spherical shape). Available Δ-XANES fingerprints for O and OH adsorption on such Pt/C materials indicate the presence of negative contribution to the edge region at high adsorbate coverages.[9,11] The effect has been attributed to increased Pt−O coordination due to the presence of subsurface oxygen at higher electrode potentials. The *all-atoms* model suggests that the effect is

caused by the change in absorption properties of adsorbate-free atoms due to a shift in the Fermi level. A greater shift is expected in the case of high coverage, while only minimal changes occur in local electronic structure for atoms that do not participate in bonding with oxygen. Due to a large number of such atoms in the realistic electrochemical cluster, the individual derivative-like $\Delta\mu$-signatures may constructively interfere with each other, which would result in the experimentally observed negative contribution.

In order to minimize the influence of the adsorbate-free atoms at the edge energy, an alignment of Fermi levels can be performed prior to subtraction of XANES if additional information about the shift is extracted from experiment data. This approach has been implemented for supported Pt catalysts in gas phase systems.[26,27] The number of Pt atoms in the supported clusters of the gas phase catalysts is comparable to the number of atoms in the Janin cluster. Thus, careful alignment procedure is of particular importance for these small clusters. Pt L_2 edge XANES of the samples with and without adsorbates were aligned under an assumption that the onset of the absorption edge is not affected by the presence of adsorbates. The Pt L_3 edge XANES were aligned relative to the L_2 edges using Extended X-ray Absorption Fine Structure (EXAFS). EXAFS refers to small oscillations in the X-ray absorption coefficient at 40−50 eV above the edge energy, which contain local structural information. The L_3 and L_2 EXAFS spectra are equivalent for the same sample under the same experimental conditions.

Unfortunately, any alignment procedure contains an uncertainty, which might lead to misinterpretation of the experimental fingerprints. Reliability of the fingerprints at the low energy side is questionable because subtraction of XANES at these energies is most sensitive to the alignment errors. The energy scale of a conventional X-ray absorption experiment is defined by mechanical motion of an X-ray double crystal monochromator. The scale slightly fluctuates from one energy scan to another. If alignment of experimental spectra produces an error ΔE, the threshold level that sets the lowest reliable magnitude of $\Delta\mu$ can be derived as follows.

The absorption coefficient is a weak function of electrochemical environment. For the purpose of error estimation, one can assume that it is only a function of X-ray energy E. A difference between two misaligned spectra of a certain representative X-ray absorption standard with absorption coefficient $\mu(E)$

$$\Delta\mu(E, \Delta E) = \mu(E + \Delta E) - \mu(E) \quad (1)$$

can be expressed through the derivative of the absorption coefficient in a vicinity of a certain energy E_0

$$\Delta\mu(E_0, \Delta E) \approx \frac{d\mu}{dE}(E_0)\Delta E \quad (2)$$

A threshold reliability level in $\Delta\mu$ can be estimated using eq 2 for any given energy. The absorption edge energy is often assigned to the inflection point. Hence, the lowest reliable $\Delta\mu$ at the absorption edge is defined by the maximum of the derivative. For Pt L_3 edge XANES this value is about 0.3 eV^{-1}. The alignment error ΔE is affected by a number of factors such as quality of reference data used for alignment, the alignment algorithm, intrinsic broadening of X-ray absorption spectra due to finite core-hole lifetime, and instrumental resolution. A reasonable estimate of $\Delta E = 0.1$ eV is on the order of a step size of a fine energy grid for XANES data collection and about 10 times less than a spectral bandpass of a conventional double crystal X-ray monochromator. This sets a reliability threshold level of $\Delta\mu = 0.03$ at the Pt L_3 edge.

Unfortunately, the threshold level is on the same order of magnitude as the experimentally observed negative contributions to Δ-XANES fingerprints at the absorption energy edge. In order to minimize the drift of the energy scale, alternative methods for obtaining Δ-XANES fingerprints are required. One of the alternatives is to utilize an energy dispersive X-ray absorption spectrometer.[3,5,28,29] Such an instrument has no moving components and enables measurements of X-ray absorption spectra with minimal energy drift. In particular, this approach has enabled observation of femtometre-scale atomic displacements by difference EXAFS.[29] Although, a higher noise level (i.e., lower sensitivity) is expected in comparison to traditional XAS (energy scans), successful applications of the dispersive method to electrochemical systems have been reported.[3,5,30] In-situ EXAFS spectra were recorded and quantitatively analyzed. EXAFS is a weak effect in comparison to XANES. This suggests applicability of the dispersive method to Δ-XANES. It is expected that minimization of energy drifts by using alternative methods for the data collection would provide further insight into the effect of adsorbate-free atoms reported here.

Conclusions

The comparison of experimentally obtained difference XANES fingerprints with theoretical difference XANES of model clusters (signatures) is a strategy for the characterization of coverage dependent adsorption sites. A single oxygen atom adsorbed on the Pt_6 Janin cluster modifies X-ray absorption properties of the adsorbate-free Pt atoms. It has been shown how the related electronic effects affect theoretical XANES spectra. Difference XANES calculated on a per atom basis reveal contribution of the adsorbate-free atoms to the averaged Δ-XANES response of the system. The resulting *all-atoms* signatures of the bridged, fcc, and hcp configurations are different from the corresponding *limited-adsorber* signatures currently used for surface adsorbate $\Delta\mu$ analysis. The most substantial difference is a negative contribution due to a positive shift of the absorption edge for the adsorbate-free atoms, which is related to the shift of the cluster Fermi level upon adsorption of oxygen. The presented results indicate that interpretation of Δ-XANES fingerprints is model dependent. It is expected that for a much larger, realistic electrochemical cluster the limited adsorber signatures more adequately represent the case of low adsorbate coverage. The experimentally observed negative contributions at higher coverages can be attributed to the influence of the adsorbate-free atoms. In addition, it is emphasized that the experimental Δ-XANES fingerprints are subject to an alignment error due to a shift of the energy scale in an experiment. The related uncertainty impedes discrimination of the fingerprints.

Acknowledgment. Thanks are due to David E. Ramaker, John J. Rehr, Carlo U. Segre, and Eugene S. Smotkin for helpful discussions.

Supporting Information Available: Figure 1S and Table 1S. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Iwasawa, Y. *X-ray Absorption Fine Structure for Catalysts and Surfaces*; World Scientific: Singapore, New Jersey, London, Hong Kong, 1996; Vol. 2, pp 3−6.

(2) McBreen, J.; Mukerjee, S. *J. Electrochem. Soc.* **1995**, *142*, 3399.

(3) Mathew, R.; Russell, A. *Top. Catal.* **2000**, *10*, 231.

(4) Roth, C.; Martz, N.; Buhrmester, T.; Scherer, J.; Fuess, H. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3555.

(5) Allen, P. G.; Conradson, S. D.; Wilson, M. S.; Gottesfeld, S.; Raistrick, I. D.; Valerio, J.; Lovato, M. *J. Electroanal. Chem.* **1995**, *384*, 99.

(6) Stoupin, S.; Chung, E.-H.; Chattopadhyay, S.; Segre, C. U.; Smotkin, E. S. *J. Phys. Chem. B* **2006**, *110*, 9932.

(7) Stoupin, S.; Rivera, H.; Li, Z.; Segre, C. U.; Korzeniewski, C.; Casadonte, D. J., Jr.; Inoue, H.; Smotkin, E. S. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6430.

(8) Teliska, M.; O'Grady, W. E.; Ramaker, D. E. *J. Phys. Chem. B* **2004**, *108*, 2333.

(9) Teliska, M.; O'Grady, W. E.; Ramaker, D. E. *J. Phys. Chem. B* **2005**, *109*, 8076.

(10) Roth, C.; Benker, N.; Buhrmester, T.; Mazurek, M.; Loster, M.; Fuess, H.; Koningsberger, D. C.; Ramaker, D. E. *J. Am. Chem. Soc.* **2005**, *127*, 14607.

(11) Teliska, M.; Murthi, V. S.; Mukerjee, S.; Ramaker, D. E. *J. Phys. Chem. C* **2007**, *111*, 9267.

(12) Scott, F. J.; Roth, C.; Ramaker, D. E. *J. Phys. Chem. C* **2007**, *111*, 11403.

(13) Scott, F. J.; Mukerjee, S.; Ramaker, D. E. *J. Electrochem. Soc.* **2007**, *154*, A396.

(14) Oudenhuijzen, M. K.; vanBokhoven, J. A.; Miller, J. T.; Ramaker, D. E.; Koningsberger, D. C. *J. Am. Chem. Soc.* **2005**, *127*, 1530.

(15) Janin, E.; von Schenck, H.; Göthelid, M.; Karlsson, U. O.; Svensson, M. *Phys. Rev. B* **2000**, *61*, 13144.

(16) Ankudinov, A. L.; Ravel, B.; Rehr, J. J.; Conradson, S. D. *Phys. Rev. B* **1998**, *58*, 7565.

(17) Ankudinov, A. L.; Nesvizhskii, A. I.; Rehr, J. J. *Phys. Rev. B* **2003**, *67*, 115120.

(18) Ankudinov, A. L.; Rehr, J. J.; Low, J.; Bare, S. R. *Phys. Rev. Lett.* **2001**, *86*, 1642.

(19) Bazin, D.; Sayers, D.; Rehr, J. J.; Mottet, C. *J. Phys. Chem. B* **1997**, *101*, 5332.

(20) Zabinsky, S. I.; Rehr, J. J.; Ankudinov, A.; Albers, R. C.; Eller, M. J. *Phys. Rev. B* **1995**, *52*, 2995.

(21) Rehr, J. J.; Ankudinov, A.; Zabinsky, S. I. *Catal. Today* **1998**, *39*, 263.

(22) Ankudinov, A. L.; Rehr, J. J.; Low, J. J.; Bare, S. R. *J. Chem. Phys.* **2002**, *116*, 1911.

(23) Ankudinov, A. L.; Rehr, J. J.; Low, J. J.; Bare, S. R. *Top. Catal.* **2002**, *18*, 3.

(24) von Barth, U.; Grossmann, G. *Solid State Commun.* **1979**, *32*, 645.

(25) Ravel, B.; Newville, M. *Phys. Scr.* **2005**, 1007.

(26) Ramaker, D. E.; Mojet, B. L.; Garriga Oostenbrink, M. T.; Miller, J. T.; Koningsberger, D. C. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2293.

(27) Ramaker, D. E.; Teliska, M.; Zhang, Y.; Stakheev, A. Y.; Koningsberger, D. C. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4492.

(28) Pascarelli, S.; Mathon, O.; Aquilanti, G. *J. Alloys Compd.* **2004**, *362*, 33.

(29) Pettifer, R. F.; Mathon, O.; Pascarelli, S.; Cooke, M. D.; Gibbs, M. R. J. *Nature (London)* **2005**, *435*, 78.

(30) Rose, A.; South, O.; Harvey, I.; Diaz-Moreno, S.; Owen, J. R.; Russell, A. E. *Phys. Chem. Chem. Phys.* **2005**, *7*, 366.

CT800544A

JCTC Journal of Chemical Theory and Computation

Calculation of Hypershielding Contribution to Isotropic Nitrogen Shielding in Strong Magnetic Fields

Jonathan Boyd*

Department of Biochemistry, South Parks Road, Oxford, OX1 3QU, United Kingdom

Gabriel I. Pagola, Maria C. Caputo, and Marta B. Ferraro†

Departamento de Física, Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires, Ciudad Universitaria, Pab. I, (1428) Buenos Aires, Argentina

Paolo Lazzeretti

Dipartimento di Chimica dell'Università degli Studi di Modena e Reggio Emilia, Via Campi 183, 41100 Modena, Italy

Received January 18, 2009

Abstract: Hypershielding contributions to magnetic shielding of the nitrogen **N** nucleus have been evaluated for some nitroso (R**N**O) and isodiazene (R_1R_2N**N**) compounds in the presence of an external spatially uniform, time-independent magnetic field, accounting for cubic response via Rayleigh−Schrödinger perturbation theory. Numerical estimates have been obtained at the coupled Hartree−Fock level of accuracy within the conventional common-origin approach. Medium-size basis sets of gaugeless (that is, without gauge-including phase factors) Gaussian functions have been employed in a numerical test to show that the isotropic hypershielding contribution $\tau^N B^2$, $\tau^N = 1/2\langle\Sigma^N_{\alpha\beta\gamma\delta}\rangle$, eqs 2−4 in the text, to average nitrogen shielding in Ph**N**O ($\tau^N \approx 1.1 \times 10^{-5}$ ppm T^{-2}), $(CH_3)_3$C**N**O ($\tau^N \approx 2.3 \times 10^{-5}$ ppm T^{-2}), and $(CH_3)_2$N**N** ($\tau^N \approx 4.4 \times 10^{-5}$ ppm T^{-2}) are similar and quite large. For ^{15}N at the highest currently available high-resolution NMR field strength of 22.3 T ($\omega^H/2\pi = 950$ MHz, $\omega^{^{15}N}/2\pi = 96.3$ MHz) the change due to the additional shielding contribution for these compounds is between ∼0.5 and ∼2 Hz to lower frequency (upfield). Employing modern NMR instrumentation, shielding perturbations of this magnitude are, in principle, within detection limits, although instrumental instabilities and other field-dependent shielding phenomena make unambiguous detection at different field strengths difficult.

I. Introduction

Matter in the presence of intense external magnetic fields changes its basic properties and structure and leads to a multitude of interesting phenomena. As a result, strong fields are of interest in several branches of physics like atomic, molecular, or solid-state physics. A growing interest in the problem of molecules and clusters in strong magnetic fields is motivated by various technological applications in different branches of nanoscience,[1−5] and the relationship between the strength of the applied magnetic field and superconductivity has long been known.[6]

A few studies discussing effects observable in systems exposed to fields with intermediate strength (conventionally in the range from 30 to 10^4 T) are available. A review on atoms, molecules, and bulk matter has recently been published.[7] Earlier references are the conference proceedings on *Atoms and Molecules in Strong External Fields*[8] and the issue of the *International Journal of Quantum Chemistry* concerning The Properties of Molecules in Strong Magnetic Fields.[9]

* To whom correspondence should be addressed.
† Member of Carrera del Investigador del CONICET.

10.1021/ct900034d CCC: $40.75
Published on Web 04/21/2009

Experimental evidence in the lower limit of the intermediate regime, where conventional methods of perturbation theory are applicable, has been sought by some authors.[10−19] Vaara and co-workers reported theoretical investigations of nuclear magnetic shielding in closed-shell atomic systems as a function of even powers of a strong perturbing magnetic field[11] and analyzed the magnetic-field dependence of ^{59}Co nuclear magnetic shielding in Co(III) complexes.[12] A computational scheme based on Rayleigh−Schrödinger perturbation theory has been developed to evaluate 10 propagators appearing in the quantum mechanical definition of the fourth-rank hypershielding $\langle\Sigma^I_{\alpha\beta\gamma\delta}\rangle$ of the Ith nucleus in a molecule.[18]

Attempts have recently been made to evaluate fourth-rank hypermagnetisabilities in small-[13] and medium-size molecules,[15] conjugated systems, and fullerene.[16] Sum rules for invariance of the fourth-rank hypermagnetizability in a gauge translation have been examined in small molecules.[17] The effects of an intense magnetic field on the electron distribution and on the magnetizability have been studied for rare gas atoms.[14] Induced electron current densities depending on the third power of the applied field have been considered for aromatic and antiaromatic compounds, showing that the nonlinear response can cause a switch from diatropic to paratropic behavior and vice versa.[20]

Tellgren et al. implemented a nonperturbative scheme using London orbitals and applied it to boron monohydride and cyclobutadiene, showing the highly nonlinear behavior of the former, which is typical of a system turning from paramagnetic to diamagnetic and going through an energy minimum in the presence of a strong magnetic field.[21]

The investigations reported so far for small molecules indicate that hypermagnetizabilities and magnetic hypershieldings of H, C, N, and O nuclei are very small.[13,15] Therefore, nonlinear response to strong magnetic fields seems to be quite difficult to detect experimentally. On the other hand, the present study shows that the fourth-rank magnetic hypershielding for the nitrogen nucleus, $\Sigma^N_{\alpha\beta\gamma\delta}$, of some nitroso (R**N**O) and isodiazene ($R_1R_2N\mathbf{N}$) compounds could, in principle, be experimentally detectable. The theoretical approach employed in this study is described in ref 18; only a brief summary is outlined in section II. Numerical results for the hypershielding contribution to the isotropic nitrogen shielding are reported and discussed in section III.

II. Fourth-Rank Nuclear Magnetic Hypershielding

The interaction energy of a diamagnetic molecule, i.e., a closed-shell system even under time reversal, in the electronic reference state a, in the presence of a static and spatially uniform external magnetic field **B**, and of an intramolecular magnetic dipole $\mathbf{m}_I$ at the Ith nucleus with position $\mathbf{R}_I$, can be expressed as a Taylor series[18]

$$\begin{aligned}\Delta W_a &= W_a^{(2)} + W_a^{(4)} + ... \\ &= \sigma^I_{\alpha\beta} m_{I\alpha} B_\beta + \frac{1}{6}\Sigma^I_{\alpha\beta\gamma\delta} m_{I\alpha} B_\beta B_\gamma B_\delta + ... \end{aligned} \quad (1)$$

Only terms relevant for the following discussion are retained in expression 1, e.g., $\sigma^I_{\alpha\beta}$is the second-rank magnetic shielding at nucleus I,[22] and the fourth-rank tensor $\Sigma^I_{\alpha\beta\gamma\delta}$ accounts for nonlinear response in **B**. NMR relaxation phenomena in liquids are typically treated using semiclassical relaxation theory, where the imaginary part of the time-dependent quantum mechanical master equation gives rise to a small change to the energy (1) of the system and a small shielding perturbation.[23] An estimate of the magnitude of this phenomenon is presented later for a small molecule in a nonviscous liquid.

The field-dependent magnetic shielding of nucleus I in a molecule in the presence of the magnetic field **B** is defined by the expression[18]

$$\sigma^I_{\alpha\beta}(\mathbf{B}) = \sigma^I_{\alpha\beta} + \frac{1}{2}\Sigma^I_{\alpha\beta\gamma\delta} B_\gamma B_\delta + ... \quad (2)$$

The hypershielding tensor is defined as the fourth derivative of the molecular energy in the limit of vanishing perturbations, that is

$$\Sigma^I_{\alpha\beta\gamma\delta} = \frac{\partial^4 W_a^{(4)}}{\partial m_{I\alpha}\,\partial B_\beta\,\partial B_\gamma\,\partial B_\delta} \quad (3)$$

The hypershielding $\Sigma^I_{\alpha\beta\gamma\delta}$ can be expressed as a sum of 10 propagators[18]

$$\begin{aligned}\Sigma^I_{\alpha\beta\gamma\delta} = &-\{\hat{B}^n_{I\alpha}, \hat{m}_\beta, \hat{m}_\gamma, \hat{m}_\delta\}_{-3} - \\ &\{\hat{B}^n_{I\alpha}, \hat{m}_\beta, \hat{\chi}^d_{\gamma\delta}\}_{-2} - \{\hat{B}^n_{I\alpha}, \hat{m}_\gamma, \hat{\chi}^d_{\beta\delta}\}_{-2} - \{\hat{B}^n_{I\alpha}, \hat{m}_\delta, \hat{\chi}^d_{\gamma\beta}\}_{-2} + \\ &\{\hat{\sigma}^{dI}_{\alpha\beta}, \hat{m}_\gamma, \hat{m}_\delta\}_{-2} + \{\hat{\sigma}^{dI}_{\alpha\gamma}, \hat{m}_\beta, \hat{m}_\delta\}_{-2} + \{\hat{\sigma}^{dI}_{\alpha\delta}, \hat{m}_\gamma, \hat{m}_\beta\}_{-2} + \\ &\{\hat{\sigma}^{dI}_{\alpha\beta}, \hat{\chi}^d_{\gamma\delta}\}_{-1} + \{\hat{\sigma}^{dI}_{\alpha\gamma}, \hat{\chi}^d_{\beta\delta}\}_{-1} + \{\hat{\sigma}^{dI}_{\alpha\delta}, \hat{\chi}^d_{\gamma\beta}\}_{-1}, \end{aligned} \quad (4)$$

where $\hat{m}_\alpha$ and $\hat{\chi}^d_{\gamma\beta}$ are the electronic operators for the orbital magnetic dipole and for the diamagnetic contribution to the susceptibility tensor, $\hat{B}^n_{I\alpha}$ is the operator for the magnetic field of n electrons on nucleus I, and $\hat{\sigma}^{dI}_{\alpha\beta}$ denotes the operator for the diamagnetic contribution to the nuclear shielding.[18]

The sum of the terms in eq 4 is invariant in a gauge transformation of the vector potential in ideal cases, for instance, for exact eigenfunctions to a model Hamiltonian, or optimal variational eigenfunctions.[24] In actual calculations allowing for the algebraic approximation, the total hypershielding $\Sigma^I_{\alpha\beta\gamma\delta}$ varies in a gauge translation, that is, a change of coordinate system. Computed results would be origin independent only in the limit of complete basis sets: the closeness of values obtained in different coordinate systems is a measure of quality of the basis set. Nonetheless, the addenda in eq 4 are not uniquely defined in any case. They transform among themselves in a change of gauge, which makes an assessment of the physical meaning of individual terms quite problematic in general. However, the use of an ad hoc gauge origin, e.g., the nucleus in question, turns out to be practical in some cases.

$\Sigma^I_{\alpha\beta\gamma\delta}$ quantitatively accounts for small changes to magnetic flux density at the nucleus caused by the Lorentz magnetic force of the external magnetic field acting to deform (compress) the electron density to a new equilibrium configuration. The magnitude of the response functions in relationship 4 depends on inverse powers (specified by the subindices −1, −2, and −3) of the excitation energies.

For the systems considered here, $\{\hat{B}^n_{I\alpha},\hat{m}_\beta,\hat{m}_\gamma,\hat{m}_\delta\}_{-3}$ is strongly influenced by (i) the excitation energies $\hbar\omega_{ja}$ for

an electronic transition, $|a\rangle \rightarrow |j\rangle$, from the reference state to an excited state, e.g., electronic transitions involving low-lying states, such as $n_N \rightarrow \pi^*$, which often have the lowest excitation energies, and (ii) the energy differences $\hbar\omega_{jk}$ between two excited states, which becomes very small for near degeneracies. In the present case of RNO compounds the first term in relationship 4 makes the largest contributions to $\Sigma^N_{\alpha\beta\gamma\delta}$ (typically $\geq 85\%$) for the gauge origin on the reference nucleus.

In an isotropic medium, the molecule rotates freely with respect to a set of axes fixed to the laboratory's frame, where the stationary magnetic field, **B**, is usually aligned with the z direction. Therefore, since we want to relate shielding and hypershielding to intrinsic molecular properties, we must transform the tensors specified in space fixed axes, $\sigma^I_{\alpha\beta}$(lab) and $\Sigma^I_{\alpha\beta\gamma\delta}$(lab), to an axis system fixed to the molecule, by averaging over the surface of the unit sphere to yield the isotropic invariants. The isotropic shielding, $\langle\sigma^I\rangle$, of an NMR experiment is

$$\langle\sigma\rangle^I = \langle\sigma^I_{\alpha\beta}(\text{lab})\rangle + \frac{1}{2}\langle\Sigma^I_{\alpha\beta\gamma\delta}(\text{lab})\rangle\mathbf{B}^2$$

$$= \frac{1}{3}\sigma^I_{\alpha\alpha}(\text{mol}) + \frac{1}{30}[\Sigma^I_{\alpha\alpha\beta\beta}(\text{mol}) + \Sigma^I_{\alpha\beta\alpha\beta}(\text{mol}) + \Sigma^I_{\alpha\beta\beta\alpha}(\text{mol})]\mathbf{B}^2 \quad (5)$$

We note that in solution other processes lead to field-dependent shielding. Molecules with an anisotropic susceptibility tensor, $\chi_{\alpha\beta}$, are partially aligned when placed in a magnetic field, resulting in incomplete averaging of the anisotropy $\Delta\sigma^I = \sigma^I_{zz} - 1/2(\sigma^I_{xx} + \sigma^I_{yy})$ of the second-rank shielding tensor $\sigma^I_{\alpha\beta}$ of nucleus I and leading to an isotropic shielding contribution with $\mathbf{B}^2$ field dependence.[25] Assuming $\sigma^I_{\alpha\beta}$ axial symmetry ($\sigma^I_{xx} = \sigma^I_{yy}$ and $\Delta\sigma^I \equiv \sigma^I_{||} - \sigma^I_{\perp}$) and limited amplitude fast internal motion, the contribution due to partial orientation is

$$\sigma^I_{\chi}(\text{ppm T}^{-2}) = (S/15k_BT)\Delta\sigma^I\left(\frac{2}{3}\Delta\chi P_2(\cos\theta) + \frac{1}{2}\chi_R \sin^2\theta\cos 2\varphi\right) \quad (6)$$

where T is the absolute temperature and θ and ϕ are the angles between the chemical shielding anisotropy (CSA) unique axis and susceptibility tensor principal axes with $\Delta\chi = \chi_{zz} - 1/2(\chi_{xx} + \chi_{yy})$, $\chi_R = \chi_{xx} - \chi_{yy}$. $P_2(\cos\theta)$ is the Legendre polynomial of second degree, and S is the generalized order parameter, which acts to scale the degree of magnetic alignment depending upon the extent of any internal motions.[26] S varies between 0 and 1, with 1 corresponding to a static rigid model. The concept of a generalized order parameter has been introduced to NMR by Lipari and Szabo.[27,28]

In addition, small relaxation-induced field-dependent shielding occurs due to the dynamic frequency shift. The imaginary part of the spectral density, derived from the Fourier transform of the autocorrelation function, can be viewed as an oscillation; therefore, NMR relaxation processes can lead to frequency shifts; however, for a spin 1/2 nucleus, assuming the fast motion limit, these are often smaller than the line width.[29]

III. Results and Discussion

The shielding range for ^{15}N is large, and notably some systems show anti (that is, negative) shielding.[30] According to Saika and Slichter,[31] the wide range is primarily due to a change in local paramagnetic shielding, σ^{pI}, and for nitrogen this proposal has found qualitative support in correlation between isotropic shielding and $1/\Delta E$ (ΔE is the excitation energy from the ground state to low-lying excited states important in σ^{pN}, mainly $n_N \rightarrow \pi^*$ transitions).[32,33] Here, $\Sigma^N_{\alpha\beta\gamma\delta}$ is also assumed to be larger in systems with low-lying excited states. Solutions of isodiazenes (R_2NN) and monomers of nitroso species (RNO) are often colored, indicative of long-wavelength electronic absorption, and σ^{pN} in $-$N**N** and $-$NO groups is large.[33−35]

We calculated $\Sigma^N_{\alpha\beta\gamma\delta}$ for a series of nitroso compounds RNO, where the electronegativity of the R$-$ group is varied, so its inductive effect, $\mathscr{I}$, ranges from electron-withdrawing ($-\mathscr{I}$) to electron donating ($+\mathscr{I}$). The groups considered are R$-$ = F$-$, Cl$-$, Br$-$, FO$-$, HO$-$, H_2N-, C_6H_5-, H$-$, CF_3-, CH_3-, and $(CH_3)_3C-$. Molecules containing these substituents range from colorless FNO to red ClNO, brown BrNO, blue-green C_6H_5NO, blue CH_3NO, and deep blue $C(CH_3)_3NO$. $\Sigma^N_{\alpha\beta\gamma\delta}$ was also calculated for two isodiazenes, H_2N**N** and $(CH_3)_2N$**N**: solutions of persistent isodiazenes are deep purple,[34] and the terminal nitrogen exhibits large (negative) antishielding.

The computational procedure is based on the coupled Hartree−Fock (CHF) approach implemented in the DALTON package.[36] The calculations employed medium-size basis sets, referred to as I → VI, that is, I, aug-cc-pCVTZ, from refs 37 and 38 for ClNO and BrNO; II, aug-cc-pCVTZ-CTOCD-uc, (C, N, O, F: 11s11p6d2f/H: 6s5p2d) contracted to [C, N, O, F: 5s11p6d2f/H: 4s5p2d] from ref 39 and 40 for FONO, HONO, H_2NNO, H_2NN, and HNO; III, $sp_{rt}d_{td}f_2$/sp_td, (C, N, O: 12s9p5d2f/H:7s4p1d) contracted to [C, N, O: 9s9p5d2f/H: 4s4p1d] from refs 41 and 42 for HONO; IV, Huz-III for CF_3NO, C_6H_5NO, $(CH_3)_3CNO$, $(CH_3)_2NN$; V, Huz-IV for FNO, C_6H_5NO, CH_3NO, $(CH_3)_2NN$ from ref 43; VI, aug-cc-pCVQZ from ref 37 and 38 for BrNO. These basis sets were found to be large enough to obtain reasonable estimates of nitrogen hypershielding at the Hartree−Fock (HF) level of accuracy.

Simple chemical-bonding consideration would suggest that the wave function of the systems of this investigation, nitroso compounds and, even more so, isodiazenes, has a significant multireference character. As Hartree−Fock perturbation theory is likely to suffer from triplet and, possibly, singlet instabilities, random-phase approximation (RPA) singlet and triplet excitation frequencies were evaluated in a check on FNO and H_2NN via the SYSMO code.[44] We found 1908 positive singlet excitation frequencies for the former and 1936 for the latter, excluding any problem of singlet instability in the calculation of $\langle\sigma^N\rangle$ and $\langle\Sigma^N\rangle$. On the other hand, triplet instabilities were found for both systems, which however do not affect shielding and hypershielding estimates.

The molecular geometries were optimized at the density functional theory (DFT) B3LYP level[45] employing the 6-31G** basis set.[46,47]

Table 1. Nitrogen Isotropic Hypershielding for Some Nitroso R**N**O and Isodiazene RN**N** Compounds[a]

molecule	basis set	$\langle\Sigma^N_{iiii}\rangle \times 10^{-4}$ [b]	$\langle\sigma^N\rangle$ [c]	$(1/2)\langle\Sigma^N\rangle B^2$ [d]
FNO	V	3.51 (3.35)	−335.3 (−264.6) [−236[e]]	3×10^{-3} (3×10^{-3})
FONO	II	7.20	−441.0	7×10^{-3}
ClNO	I	9.99	−431.9	9×10^{-3}
HONO	II	10.09	−484.2	9×10^{-3}
	III	10.16	−485.6	9×10^{-3}
H_2NNO	II	10.29	−531.5	9×10^{-3}
BrNO	I	17.80	−384.1	0.02
	VI	21.19	−441.9	0.02
C_6H_5NO	IV	125.5 (65.8)	−1067.2 (−631.7)	0.11 (0.06)
	V	120.3 (64.0)	−1063.0 (−629.9) [−667[f]]	0.11 (0.06)
CF_3NO	IV	145.9	−431.9	0.13
CH_3NO	V	203.6	−1269.3	0.18
HNO	II	229.8	−1240.5	0.21
$C(CH_3)_3NO$	IV	248.4 (87.4)	−1373.2 (−791.1) [−735[g]]	0.23 (0.07)
H_2NN	II	423.1	−2025.8	0.38
$(CH_3)_2NN$	IV	482.3	−2110.0	0.44
	V	437.0	−2064.7	0.40

[a] Hartree−Fock results. The gauge origin is taken at the nucleus in question. Corresponding values via DFT- Keal-Tozer KT3 functional[48,49] are given in parentheses, and experimental values are reported in brackets. [b] In SI atomic units. The conversion factor to SI units is $1.80997698 \times 10^{-11}$ T^{-2}, using the CODATA values of the fundamental constants 2002.[56] [c] In ppm. [d] Isotropic contribution to $\langle\sigma^N\rangle$, eqs 2 and 5, in ppm, for $B = 100$ T. [e] Estimated from the average ^{15}N chemical shift, 128 ppm, of FNO with respect to aqueous nitrite, ref 57, the ^{15}N chemical shift of the aqueous nitrite ion, 608 ppm, with respect to anhydrous NH_3, ref 33, and the absolute shielding, 244.6 ppm, of nitrogen in NH_3, ref 58. [f] Estimated from the ^{15}N shielding, −531.52 ppm, from ref 59 (sample in acetone and referenced to CH_3NO_2); converted to the shielding scale assuming σ^N_{av} (CH_3NO_2) = −135.8 ppm from ref 58. The experimental ^{15}N shielding measured in this work (sample in acetone) $\sigma^N_{av} = -666.4$ ppm. [g] Estimated from the ^{15}N shielding, −599.40 ppm, from ref 35 (sample in acetone and referenced to CH_3NO_2); converted to the shielding scale assuming σ^N_{av} (CH_3NO_2) = −135.8 ppm from ref 58.

The calculated isotropic contribution from hypershielding $\langle\Sigma^N\rangle$ to isotropic shielding $\langle\sigma^N\rangle$ will be origin independent only for complete basis sets; see the previous discussion. However, for $(CH_3)_2N\mathbf{N}$ and HO**N**O, it was found for two different origins of the vector potential, namely, the nucleus in question and the center of mass (CM), hypershielding differs by less than 2×10^{-9} ppm T^{-2} with basis sets in Table 1. Therefore, relying on that criterion for these types of molecular systems, we conclude that basis sets I−VI are suitable for quantitative estimates of nitrogen hypershielding.

Although we do not claim that the calculated second- and fourth-rank shielding tensors of the N nucleus have converged to the HF limit, the values reported in Table 1 for HONO, C_6H_5NO, and $(CH_3)_2NN$ for two basis sets document the quality of the present predictions. However, we also note that for these types of molecular systems the HF approach, due to lack of electron correlation, often provides quite poor values for $\langle\sigma^N\rangle$. Very much better agreement with experimental isotropic shielding data is achieved via the DFT approach, allowing for the Keal and Tozer KT3 functional,[48,49] as implemented in the DALTON package.[36] The KT3 functional was found to be the most efficient for calculating shielding constants in a series of small molecules[50] and for studies of hydrogen bonding.[51] Also, in the present research the computed isotropic shielding values for FNO, C_6H_5NO and $(CH_3)_3CNO$ are each quite close to experiment.

An indication obtained here is that electron-correlation effects act to lower the magnitude of $\langle\sigma^N\rangle$ and $\langle\Sigma^N\rangle$. On the other hand, DFT methods employed in this work are not designed to properly deal with nondynamical correlation. Therefore, further work will be required to assess the importance of electron correlation in nuclear hypershielding of nitrogen in RNO and R_1R_2NN compounds; however, this is beyond the scope of the present paper, whose primary concern is to identify molecular systems as potential candidates for experimental detection.

Clearly for this series of nitroso derivatives the magnitude of nitrogen hypershielding is sensitive to the inductive character of R−; as the electronegativity of R increases, hypershielding is reduced, whereas electron-donating groups such as $(CH_3)_3C-$ markedly increase hypershielding. The model isodiazenes show similar large hypershielding effects with $\tau^N \equiv 1/2\langle\Sigma^N\rangle \approx 4.4 \times 10^{-5}$ ppm T^{-2}.

The dominant contribution to hypershielding arises from the response function $\{\hat{B}^n_{l\alpha},\hat{m}_\beta,\hat{m}_\gamma,\hat{m}_\delta\}^{mol}_{-3}$; see the previous discussion. This term is positive for the systems studied assuming the origin of the gauge on the N nucleus and, therefore, acts to *increase* isotropic shielding.

For PhNO, $\tau^N \approx 1.1 \times 10^{-5}$ ppm T^{-2} and assuming spectrometer magnetic field strengths for ^{15}N of 8.45 T ($\omega^N/2\pi$ = 36.5 MHz) and 22.3T ($\omega^N/2\pi$ = 96.3 MHz), the shielding change due to hypershielding as the field increases from 8.45 T to 22.3 T is $\approx 4.7 \times 10^{-3}$ ppm (≈0.5 Hz) upfield, i.e., to lower resonance frequency. In principle, although small, a shielding change of this magnitude is detectable as modern high-resolution NMR spectrometers are capable of resolving differences of ∼0.15 Hz, as occurs for the H resonance in $CHCl_3$ due to[35,37] Cl isotope shifts.[52]

However, several experimental factors, including sample stability (over periods of many days) and the requirement for identical sample temperature on spectrometers operating with different field strengths, have made it difficult to achieve the required experimental accuracy and reproducibility to adequately test the predictions. The measured ^{15}N temperature coefficient, $\Delta\langle\sigma^N\rangle/\Delta T$, of $Ph^{15}NO$ in acetone-d_6 over the small range 20−21 °C is ca. -2.6×10^{-3} ppm per 0.1 °C. Furthermore, in solution NMR due to the effect of susceptibility anisotropy, PhNO is expected to be weakly oriented

by the static magnetic field. A calculation of $\Delta\chi$, χ_R, and $\Delta\sigma^N$ for PhNO gave -1.94×10^{-27} J T^{-2}, 1.05×10^{-27} J T^{-2}, and -1685 ppm, respectively, with $\theta = \pi/2$ and $\phi = 4.3°$. Assuming $T = 293$ K and limited fast internal motion, $S = 0.8$, sample orientation could lead to an isotropic field-dependent shielding contribution to σ_χ^N as big as -2.6×10^{-5} ppm T^{-2}.

The magnitude is similar but opposite in sign to the hypershielding effect. The calculated CSA, $\Delta\sigma^N$, for PhNO is large; however, the shielding tensor characteristics are similar to experimental data from *p*-nitroso-*N*,*N*-dimethylalanine.[53] Furthermore, the average shielding estimated for the nitrogen nucleus via some preliminary DFT calculations,[54] using the KT3 scheme functional[48,49] available in the DALTON code,[36] -629.9 ppm, is in good agreement with experiment, -666 ppm (sample in acetone).

Assuming the rotational diffusion characteristics of PhNO are similar to toluene[55] with symmetric top rotational diffusion and $D_{||} = 3.8 \times 10^{10}$ s^{-1}, $D_\perp = 9.1 \times 10^{10}$ s^{-1}, $1/\tau_{2n} = 6D_\perp + n^2(D_{||} - D_\perp)$, and $\omega^N\tau_{2n} \ll 1$, the calculated dynamic frequency shift from autocorrelated ^{15}N−H dipolar and ^{15}N CSA relaxation is small ($<10^{-8}$ ppm T^{-2}) and field-dependent shielding contribution from the dynamic frequency shift can be neglected for PhNO.

IV. Concluding Remarks

Measurement of response properties of a molecule in a strong magnetic field, such as susceptibility and nuclear shielding, as a function of the field strengths currently attainable in laboratory, have so far been considered quite difficult.

Theoretical results obtained in this work at the Hartree−Fock and DFT levels of accuracy, within the common origin approach for the vector potential, indicate that the hypershielding of nitrogen nuclei in molecules containing nitroso and diazene moieties might be experimentally observed. Therefore, the recent introduction of NMR instruments using superconducting magnets with strong flux density increases the possibility that useful information may soon become available from the field dependence of spectral parameters such as chemical shielding of nitrogen in R**N**O and R_1R_2N**N** compounds.

The increased chemical shieldings (upfield) arising from quadratic dependence of the nuclear shielding constant in $(CH_3)_3$CNO and $(CH_3)_2$NN estimated via Hartree−Fock calculations are, respectively, ∼0.2 and ∼0.4 ppm for a field of 100 T, which supports the expectation that hypershielding effects in NMR spectra are, in principle, within detection limits, although other field-dependent processes, such as susceptibility-induced partial orientation effects, may act to reduce the magnitude of the hypershielding phenomenon as predicted here for C_6H_5NO. Electron correlation effects partially investigated here are quite sizable and should be considered for accurate quantitative predictions. However, further theoretical and experimental investigation of quadratic dependence of NMR parameters does now, at least, continue to merit consideration.

Acknowledgment. Financial support to this work from the Italian MURST (Ministero dell'Università e della Ricerca Scientifica e Tecnologica), via PRIN funds, from the University of Buenos Aires (UBACYT X-079), and from the Argentinian CONICET is gratefully acknowledged.

References

(1) Bachtold, A.; Strunk, C.; Salvetat, J.-P.; Bonard, J.-M.; Forró, L.; Nussbaumer, T.; Schönberger, C. Aharonov-Bohm oscillations in carbon nanotubes. *Nature (London)* **1999**, *397*, 673.

(2) Zaric, S.; Ostojic, G. N.; Kono, J.; Shaver, J.; Moore, V. C.; Strano, M. S.; Hauge, R. H.; Smalley, R. E.; Wei, X. Optical Signatures of the Aharonov-Bohm Phase in Single-Walled Carbon Nanotubes. *Science* **2004**, *304*, 1129.

(3) Coskun, U. C.; Wei, T.-C.; Vishveshwara, S.; Goldbart, P. M.; Bezryadin, A. *h/e* Magnetic Flux Modulation of the Energy Gap in Nanotube Quantum Dots. *Science* **2004**, *304*, 1132.

(4) Compernolle, S.; Chibotaru, L. F.; Ceulemans, A. Novel type of magnetic response in carbon nanomaterials. *Chem. Phys. Lett.* **2006**, *428*, 119.

(5) Compernolle, S.; Chibotaru, L. F.; Ceulemans, A. Vortices and their relation to ring currents and magnetic moments in nanographenes in high magnetic field. *J. Chem. Phys. B* **2006**, *110*, 19340.

(6) Rasolt, M.; Tešanović, Z. Theoretical aspects of superconductivity in very high magnetic fields. *Rev. Mod. Phys.* **1992**, *64*, 709.

(7) Lai, D. Matter in strong magnetic fields. *Rev. Mod. Phys.* **2001**, *73*, 629.

(8) Proceedings of the 172nd WE-Heraeus-Seminar on Atoms and Molecules in Strong External Fields, Bad Honnef, Germany, April 7−11, 1997. In *Atoms and Molecules in Strong External Fields*, 2nd ed.; Schmelcher, P., Schweizer, W., Eds.; Springer: New York, USA, 1998.

(9) Runge, K.; Sabin, J. R. Introduction to the Workshop on Properties of Molecules in Strong Magnetic Fields. *Int. J. Quantum Chem.* **1997**, *64*, 495, and references therein.

(10) Ramsey, N. F. Possibility of Field-Dependent Nuclear Magnetic Shielding. *Phys. Rev. A* **1970**, *1*, 1320.

(11) Vaara, J.; Manninen, P.; Lounila, J. Magnetic Field-Dependence of Nuclear Magnetic Shielding in Closed-Shell Atomic Systems. *Chem. Phys. Lett.* **2003**, *372*, 750.

(12) Manninen, P.; Vaara, J. Magnetic-field dependence of ^{59}Co nuclear magnetic shielding in Co(III) complexes. *Phys. Rev. A* **2004**, *69*, 022503.

(13) Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. Calculation of the fourth-rank hypermagnetizability of some small molecules. *J. Chem. Phys.* **2004**, *120*, 9556.

(14) Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. Non-linear ring currents: effect of strong magnetic fields on π-electron circulation. *Chem. Phys. Lett.* **2004**, *400*, 133.

(15) Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. Nonlinear response of the benzene molecule to strong magnetic fields. *J. Chem. Phys.* **2005**, *112*, 074318.

(16) Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. Fourth-rank hypermagnetizability of medium-size planar conjugated molecules and fullerene. *Phys. Rev. A* **2005**, *72*, 033401:1.

(17) Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. Sum rules for invariance of the fourth-rank hypermagnetisability in a gauge translation. *Chem. Phys. Lett.* **2005**, *408*, 403.

(18) Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. Calculation of the fourth-rank nuclear magnetic hypershielding of some small molecules. *Phys. Rev. A* **2006**, *74*, 022509.

(19) Žaucer, M.; Amńan, A. Magnetic Field-Dependent Molecular Susceptibility. *Phys. Rev. A* **1977**, *16*, 475.

(20) Soncini, A.; Fowler, P. W. Non-linear ring currents: effect of strong magnetic fields on π-electron circulation. *Chem. Phys. Lett.* **2004**, *400*, 213.

(21) Tellgren, E. I.; Soncini, A.; Helgaker, T. Nonperturbative ab initio calculations in strong magnetic fields using London orbitals. *J. Chem. Phys.* **2008**, *129*, 154114.

(22) Ramsey, N. F. Magnetic Shielding of Nuclei in Molecules. *Phys. Rev.* **1950**, *78*, 699.

(23) Abragam, A. *The Principles of Nuclear Magnetism*; Oxford: Clarendon Press: London, 1961; p 279.

(24) Epstein S. T. *The Variation Method in Quantum Chemistry*; California University, Academic Press: New York, 1974.

(25) Bothner-By, A. A. Magnetic Field Induced Alignment of Molecules. In *Encyclopedia of Nuclear Magnetic Resonance*; Grant, D. M., Harris, R. K., Eds.; John Wiley & Sons: Chichester, 1995; pp 2932−2938.

(26) Ottiger, M.; Tjandra, N.; Bax, A. Magnetic Field Dependent Amide ^{15}N Chemical Shifts in a Protein-DNA Complex Resulting from Magnetic Ordering in Solution. *J. Am. Chem. Soc.* **1997**, *119*, 9825.

(27) Lipari, G.; Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance. *J. Am. Chem. Soc.* **1982**, *104*, 4546.

(28) Lipari, G.; Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J. Am. Chem. Soc.* **1982**, *104*, 4559.

(29) Werbelow, L. G. Dynamic Frequency Shift. In *Encyclopedia of Nuclear Magnetic Resonance*; Grant, D. M., Harris, R. K., Eds.; John Wiley & Sons: Chichester, 1995; pp 1776−1783.

(30) Baker, M. R.; Anderson, C. H.; Ramsey, N. F. Nuclear Magnetic Antishielding of Nuclei in Molecules. Magnetic Moments of F19, N14, and N15. *Phys. Rev.* **1964**, *133*, A1533.

(31) Saika, A.; Slichter, C. P. A Note on the Fluorine Resonance Shifts. *J. Chem. Phys.* **1954**, *22*, 26.

(32) Herbison-Evans, D.; Richards, R. E. ^{14}N chemical shifts in organic compounds. *Mol. Phys.* **1964**, *8*, 19.

(33) Lambert, J. B.; Roberts, J. D. Nitrogen-15 Magnetic Resonance Spectroscopy. V. Oxygen-Nitrogen Compounds. *J. Am. Chem. Soc.* **1965**, *87*, 4087.

(34) Dervan, P. B.; Squillacote, M. E.; Lahti, P. M.; Sylwester, A. P.; Roberts, J. D. Nitrogen-15 NMR spectrum of a 1,1-diazene. N-(2,2,6,6-tetramethylpiperidyl)nitrene. *J. Am. Chem. Soc.* **1981**, *103*, 1120.

(35) Witanowski, M.; Biedrzycka, Z.; Webb, G. A. Solvent Effects on the Nitrogen NMR Shielding of 2-Methyl-2-nitrosopropane and its Azodioxy Dimer. *Magn. Reson. Chem.* **1996**, *34*, 233.

(36) *DALTON, An electronic structure program*, Release 2.0; http://www.kjemi.uio.no/software/dalton/ (accessed Feb 24, 2006), 2005.

(37) Dunning, T. H., Jr. Gaussian Basis Set for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.

(38) Woon, D. E.; Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. V. Core-ion valence basis sets for boron through neon. *J. Chem. Phys.* **1995**, *103*, 4572.

(39) Ligabue, A.; Sauer, S. P. A.; Lazzeretti, P. Correlated and gauge invariant calculations of nuclear magnetic shielding constants using the continuous transformation of the origin of the current density approach. *J. Chem. Phys.* **2003**, *118*, 6830.

(40) The aug-cc-pCVTZ/CTOCD-uc and $sp_td_{td}f_2/sp_td$ basis sets can be downlowed from http://fyskem.ki.ku.dk/sauer/BasisSets.

(41) Sauer, S. P. A.; Paidarová, I.; Oddershede, J. Correlated and Gauge Origin Independent Calculations of Magnetic Properties. II. Shielding Constants of Simple Singly Bonded Molecules. *Theor. Chim. Acta* **1994**, *88*, 351.

(42) Sauer, S. P. A.; Paidarová, I.; Oddershede, J. Correlated and Gauge Origin Independent Calculations of Magnetic Properties. I. Triply Bonded Molecules. *Mol. Phys.* **1994**, *81*, 87.

(43) Huzinaga, S.; Andzelm, J.; Klobukowsi, M.; Radzio-Andzelm, E.; Sakai, Y.; Tatewaki H. *Gaussian Basis Sets for Molecular Calculations*; Elsevier: Amsterdam, 1984.

(44) Lazzeretti, P.; Malagoli, M.;Zanasi, R. Technical report on project "Sistemi informatici e calcolo parallelo", Research Report 1/67, CNR, 1991.

(45) Becke, A. D. Density-functional termochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.

(46) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *J. Chem. Phys.* **1988**, *89*, 2193.

(47) Petersson, G. A.; Tensfeldt, T. G.; Montgomery, J. A., Jr. A complete basis set model chemistry. III. The complete basis set-quadratic configuration interaction family of methods. *J. Chem. Phys.* **1991**, *94*, 6091.

(48) Keal, W.; Tozer, D. J. The exchange-correlation potential in Kohn-Sham nuclear magnetic resonance shielding calculations. *J. Chem. Phys.* **2003**, *119*, 3015.

(49) Keal, W.; Tozer, D. J. A semiempirical generalized gradient approximation exchange-correlation functional. *J. Chem. Phys.* **2004**, *121*, 5654.

(50) Ligabue, A.; Sauer, S. P. A.; Lazzeretti, P. Gauge invariant calculations of nuclear magnetic shielding constants using the continuous transformation of the origin of the current density approach. II. Density functional and coupled cluster theory. *J. Chem. Phys.* **2007**, *126*, 154111.

(51) Kongsted, J.; Aidas, K.; Mikkelsen, K. V.; Sauer, S. P. A. On the accuracy of density functional theory to predict shifts in nuclear magnetic resonance shielding constants due to hydrogen bonding. *J. Chem. Theor. Comput.* **2008**, *4*, 267.

(52) Soffe, N.; Boyd, J.; Leonard, M. The construction of a high-resolution 750 MHz probehead. *J. Magn. Reson. A* **1995**, *16*, 117.

(53) Lumsden, M. D.; Wu, G.; Wasylishen, R. E.; Curtis, R. D. Solid-state nitrogen-15 NMR studies of the nitroso group in the nitrosobenzene dimer and p-nitroso-N, N-dimethylaniline. *J. Am. Chem. Soc.* **1993**, *115*, 2825.

(54) Boyd, J. Pagola, G. I.; Caputo, M. C.; Ferraro, M. B.; Lazzeretti, P. To be published.

(55) Sturz, L.; Dölle, A. Anisotropic Reorientational Dynamics of Toluene in Neat Liquid. A ^{13}C Nuclear Magnetic Relaxation Study. *J. Phys. Chem. A* **2001**, *105*, 5055.

(56) Mohr, P. J.; Taylor, B. N. CODATA Recommended Values of the Fundamental Physical Constants: 2002. *Rev. Mod. Phys.* **2005**, *77*, 1.

(57) Andersson, L. O.; Mason, J. B.; van Bronswijk, W. Nitrogen Nuclear Magnetic Resonance. Part 1. The Nitroso(Nitrosyl) Group. *J. Chem. Soc. A* **1970**, *1970*, 296.

(58) Jameson, C. J.; Jameson, A. K.; Oppusunggu, D.; Wille, S.; Burrell, P. M.; Mason, J. ^{15}N Nuclear Magnetic Shielding Scale from Gas Phase Studies. *J. Chem. Phys.* **1981**, *74*, 81.

(59) Witanowski, M.; Biedrzycka, Z.; Sicinska, W.; Webb, G. A. Solvent-Induced Effects on the Nitrogen NMR Shieldings of Some Nitrosobenzene Systems. *Magn. Reson. Chem.* **1997**, *35*, 262.

CT900034D

JCTC Journal of Chemical Theory and Computation

Adsorption of Ethylene, Vinyl, Acetic Acid, and Acetate Species on PdAu(111) and PdAu(100) Surface Alloys: A Cluster Model Study

Ivan Rivalta, Gloria Mazzone, Nino Russo, and Emilia Sicilia*

Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni Parallele e Distribuite-Centro d'Eccellenza MURST, Università della Calabria, I-87030 Arcavacata di Rende, Italy

Received January 7, 2009

Abstract: The adsorption properties on PdAu surface alloys of ethylene and acetic acid molecules along with their derived vinyl and acetate surface species have been investigated by density functional theory calculations. Large clusters have been used to model second-neighbor Pd monomer pair ensembles on PdAu(100) and PdAu(111) surface alloys. Ethylene and acetic acid are weakly bonded to the Pd monomers, while vinyl and acetate are strongly bonded to both Pd and Au atoms being very stable surface species. The ligand effect of the gold atoms surrounding the Pd monomers has been shown to be stronger in the more dense PdAu(111) surface alloy. Cluster model results are in good agreement with experimental evidence providing important insight on the adsorption bonding modes, the assignment of the infrared features, and the preferred adsorption sites.

1. Introduction

Bimetallic systems constitute a broad class of selective catalysts that attempt to exploit and combine the different chemical properties of various metals toward a given chemical reaction.[1−4] Among the several binary systems that are often used in basic and applied research, PdAu provides an example of a rather versatile catalytic system. Mixtures of Pd and Au are frequently used as catalysts for a number of reactions (e.g., CO oxidation, cyclotrimerization of acetylene, synthesis of vinyl acetate monomer, selective oxidation of alcohols to aldehydes)[5−8] and applications including hydrogen fuel cells and pollution control.[9,10] It has been shown that the addition of Au to Pd significantly enhances the catalytic activity, selectivity, and stability of Pd-based catalysts; therefore, Pd−Au surface alloys have been the subject of many investigations.[6,7,11−18]

One of the major uses of gold/palladium catalysis is for vinyl acetate monomer (VAM) production. Indeed, the synthesis of vinyl acetate monomer from ethylene, acetic acid, and oxygen is catalyzed by both supported palladium and palladium−gold alloys, but alloying with gold leads to a substantial increase in selectivity.[19] Although this process is a mature industrial reaction, the nature of the key reaction intermediates and mechanism as well as the promotional role of Au are still unresolved questions. Two different surface reaction mechanisms, both invoking coupling of acetate and ethylenic species as the rate-limiting step, have been proposed for VAM synthesis. The one proposed by Samanos et al. involves the coupling of ethylene with adsorbed acetate to form, as intermediate, ethyl acetate that undergoes β-hydride elimination to form VAM (Samanos-type mechanism).[20] Alternatively, as suggested by both Moiseev[21,22] and Nakamura and Yasui,[23] ethylene could adsorb and subsequently dehydrogenate to form a vinyl species that couples with the coadsorbed acetate to give VAM directly (Moiseev-type mechanism). Although support to the pathway proposed by Samanos et al. comes from the experimental results by Tysoe et al. on Pd(111) surface,[24] the reaction mechanism for VAM synthesis remains essentially uncertain.

Goodman and co-workers[25] recently demonstrated that the critical ensemble for VAM formation on a PdAu(100) surface alloy is a pair of noncontiguous, suitably spaced, Pd monomers. They have also used ultra-high-vacuum (UHV)

* To whom correspondence should be addressed. E-mail: siciliae@unical.it.

10.1021/ct9000137 CCC: $40.75
Published on Web 04/20/2009

scanning tunnelling microscopy (STM) to image the Pd monomer pairs on AuPd(100) bulk alloy surface and presented a method for increasing the concentration of these active sites.[26]

The formation of isolated Pd monomer sites influences the properties of adsorbed reagents (ethylene, acetic acid, and their derived surface species) as well. Indeed, temperature-programmed desorption (TPD) experiments have shown that ethylene binds significantly less strongly to a Pd monomer compared to a site containing contiguous Pd atoms.[27,28] On contiguous Pd sites a di-σ-bonded ethylene species and an ethylidyne species have been observed,[29] which decompose leading to the formation of carbon, while on Pd monomers surrounded by gold atoms a π-ethylene surface species has been observed.[25,28,29] The addition of Au inhibits the decomposition of ethylene, consequently forming less carbon and extending the activity of the catalyst.[25] The same effect of gold addition to Pd surfaces has been demonstrated for the acetic acid and acetate decompositions.[30,31] High-resolution electron energy loss spectroscopy (HREELS) experiments[25,29] have shown the presence of a bidentate acetate species on isolated and contiguous Pd sites on PdAu(111) surfaces at high temperatures (350 K). By use of reflection−absorption infrared spectroscopy (RAIRS),[31] intact acetic acid molecules that form catemers have been detected for Au-rich PdAu(111) surface, whereas both bidentate and monodentate acetate species have been observed on surface alloys containing small mole fractions of gold. Recently, the effect on the adsorption of reactants (in VAM synthesis) has been demonstrated to be an important factor in determining the structure sensitivity of PdAu surface alloys.[32]

Since accurate knowledge of the binding sites and of the orientation of the molecules adsorbed on surfaces is an important component of any effort to describe heterogeneously catalyzed reactions with mechanistic detail, here we present the outcome of a density functional cluster model study of the adsorption properties of ethylene, vinyl, acetic acid, and acetate species on the PdAu(111) and PdAu(100) surface alloys. Ethylene and acetic acid (along with molecular oxygen) are the starting reagents in the VAM synthesis and the species that deposit from the gas phase on the catalyst surface. Their derived species, acetate and vinyl, are the most relevant surface species if both Samanos-type and Moiseev-type mechanisms are considered viable on PdAu surfaces.

The focus of the present paper is on the local features of the interaction between adsorbed molecules and PdAu surface alloys. Consequently, we adopted for our investigation the cluster model approach as more suitable, with respect to alternative slab models, to capture the essential features of the chemisorption bond and to predict local properties such as adsorption geometries and vibrational frequencies at low coverages. In particular, cluster models have been built up to represent the second-neighbor Pd pair ensembles on both surfaces given that these are the critical ensembles in the synthesis of vinyl-acetate monomer catalyzed by PdAu alloys and supposed to be crucial for the different catalytic activities of PdAu(100) and PdAu(111) surfaces.

Results have been compared with previous periodic slab supercell calculations for the PdAu(001) surface[33,34] and also with helpful information concerning the adsorption properties of acetic acid-derived species and ethylene on Au/Pd(111) alloys as a function of temperature and alloy composition comprised in a recent series of systematic experimental studies carried out by Tysoe's group.[11,28,31]

2. Methodology and Computational Details

In order to model the second-neighbor Pd pair ensembles on PdAu(100) and PdAu(111) surfaces we have built up cluster models of large size having 57 and 64 metal atoms, respectively, in order to prevent any indirect insaturation effect for the surface atoms directly involved in the adsorbate−surface bonds. A schematic representation of the Pd_2Au_n clusters used in this work is reported in Figure 1.

Since these models have been successfully adopted for the study of CO adsorption on the PdAu surface alloys, a more detailed description of the Pd_2Au_{55}(24,21,12) and Pd_2Au_{62}(30,21,13) cluster models used here along with the cluster size dependence of the chemisorption properties can be found elsewhere.[35] The employed clusters are three-layer models centered in the active site that contains two noncontiguous Pd atoms and two Au atoms bonded to both of them. The atomic positions of the clusters have been assigned, initially, as in bulk gold, with the experimental lattice parameter of 4.08 Å. For the adsorption studies, the positions of the molecules, adsorbed on these cluster models, have been fully relaxed without symmetry constraints. Cluster atoms involved in the adsorption bonds and their first neighbors in the first and second layers have been also relaxed. The calculations have been performed with the Turbomole package[36] at the DF level, by use of the BP86 functional[37,38] within the resolution of the identity approximation for computing the electronic Coulomb interaction (RI-J).[39,40] The Stuttgart effective core potential[41] has been used to model the scalar relativistic effects replacing the 28 and 60 core electrons of palladium and gold atoms, respectively. The valence electrons of metal atoms, 18 for Pd and 19 for Au, and all electrons for C, O, and H atoms have been explicitly considered by use of the Turbomole's TZVP basis set[42] along with the corresponding TZVP auxiliary basis set.[40] The electronic ground states of bare PdAu clusters and adsorbate/cluster models (including ionic adsorbates) have been checked by studying different spin multiplicities. The lowest energy states have always shown the smallest multiplicity values (i.e., singlet or doublet states). The total energies calculated after geometry optimization have been corrected for basis set superposition error (BSSE) by use of Boys−Bernardi counterpoise calculations[43] for all adsorbate/cluster models. This correction is less than 4.0 kcal/mol per bond for the strong chemisorption of acetate radical and vinyl species on the cluster models, whereas is around 1.0−1.3 kcal/mol for the weak adsorption of ethylene and acetic acid molecules. As expected, for the absorption of charged species, i.e., CH_3COO anion, the correction is larger up to ∼7−8 kcal/mol. All the vibrational frequencies reported in this work have been corrected by use of the scaling factor (1.004) suggested for the BP86 functional in combination

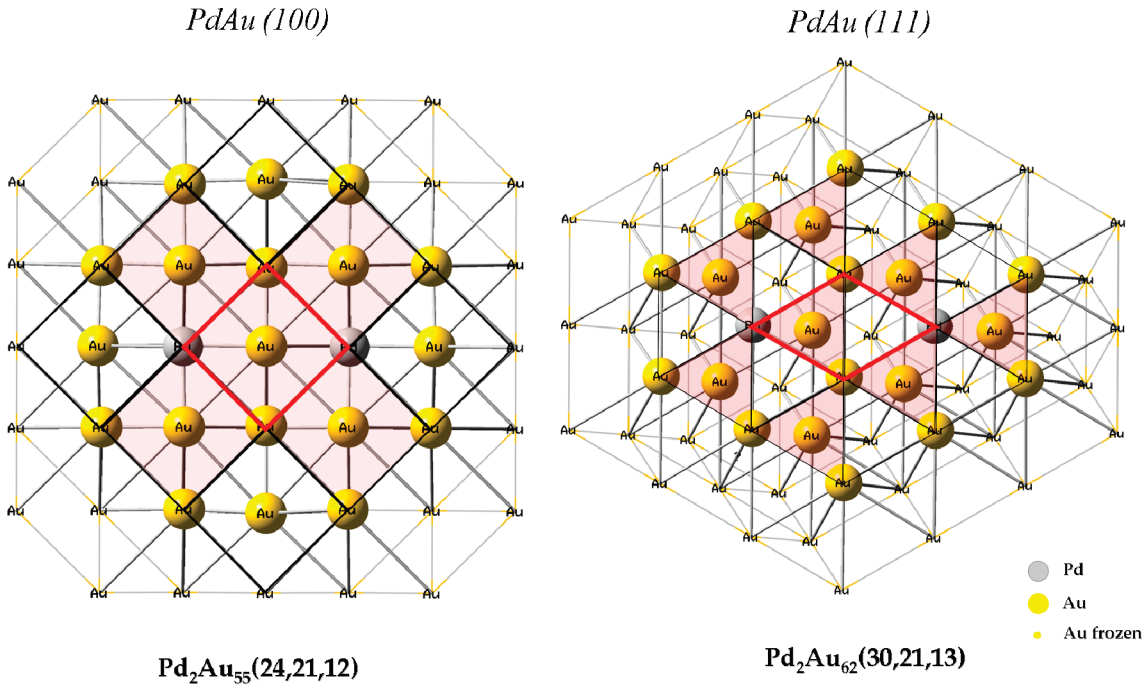


Figure 1. Schematic representation of the Pd_2Au_n clusters used in this work to model the second-neighbor Pd pair ensemble on the PdAu(100) and PdAu(111) surface alloys. Frozen Au atoms are shown as spots, while Pd and Au atoms in the first and second layer that are allowed to relax are depicted as spheres. Four-fold and three-fold (hcp) hollow sites associated with the relaxed atoms in the first layer are evidenced by planes (in red). Unit cells of the Au(100) and Au(111) pure surfaces are depicted in red.

with the TZVP basis set.[44] The atom-projected density of states (DOS) has been calculated (as implemented in Turbomole) by broadening the discrete energy levels of each cluster model with Gaussians having a width of 0.003 au and superimposing them.

3. Results and Discussion

3.1. Adsorption of Ethylene. Experimental studies by both Goodman's[25,29] and Tysoe's[28] groups have demonstrated that on gold−palladium alloys at low palladium coverages only π-bonded adsorbed ethylene on isolated Pd atoms, completely surrounded by gold, is present as surface species.

In accordance with this experimental evidence, the DF cluster model calculations performed in this study have shown that there are no stable adsorption states for a di-σ-bonded ethylene on second-neighbor Pd monomer pairs. Indeed, despite the numerous attempts to find on both PdAu(100) and PdAu(111) surface alloys a stable adsorption geometry having Pd−C and Au−C σ bonds, the ethylene molecule always moves to an atop configuration (π-bonded-type) on a Pd atom completely surrounded by gold.

The energetic, structural, and vibrational properties calculated for the π-bonded ethylene molecule adsorbed on top of Pd monomer are reported in Table 1 for both PdAu(100) and PdAu(111) palladium−gold surface alloys.

For the PdAu(100) surface, the calculated binding energy (BE) of ethylene is 13.0 kcal/mol, in good agreement with the values obtained by use of a plane-wave-based method[33] for isolated Pd monomers (12.9 kcal/mol) and second-neighbor Pd monomer pairs (15.2 kcal/mol). The geometry of the adsorbed ethylene molecule with the C−C bond along the [011] axis of the PdAu(100) surface is energetically preferred by only 0.7 kcal/mol with respect to that along the [010] axis, analogously to what has been obtained with the periodic slab supercell approach.[33]

Since the experimental BE of the π-bonded ethylene on the Pd(100) clean surface is ~19 kcal/mol[45] and this value

Table 1. BSSE-Corrected Binding Energies (BE, in kcal/mol), Equilibrium Bond Distances (*d*, in Angstroms), Angles (in degrees), and Scaled Vibrational Frequencies (in cm^{-1}) of the π-Bonded Ethylene Molecule on the Pd_2Au_{55}(24,21,12) and Pd_2Au_{62}(30,21,13) Cluster Models[a]

	PdAu(100)	PdAu(111)
BE	13.0	5.5
d(C−C)	1.383	1.374
d(Pd−C)	2.299	2.383
θ(CCH)	120.6	121.1
ϕ_{tilt}	8.4	5.1
ν(C−H)	3184 (w)	3178 (w)
	3159 (w)	3152 (w)
	3083 (w)	3083 (w)
	3080 (w)	3076 (w)
ν(C−C) + $\delta(CH_2)$	1530 (w)	1607 (w)
	1422 (w)	1430 (w)
	1276 (w)	1336 (w)
	1203 (w)	1206 (w)
$\omega(CH_2)$	958 (w)	1005 (w)
	922 (s)	938 (vs)
	900 (s)	925 (s)
	818 (w)	811 (w)
ν(Pd−C)	175 (w)	158 (w)

[a] The ϕ_{tilt} angle indicates the tilting angle between the ethylene molecular plane and the surface plane. ν stretching; δ in-plane bending, scissoring and rocking; ω out of plane bending, twisting and wagging. (vs) very strong; (s) strong; (w) weak.

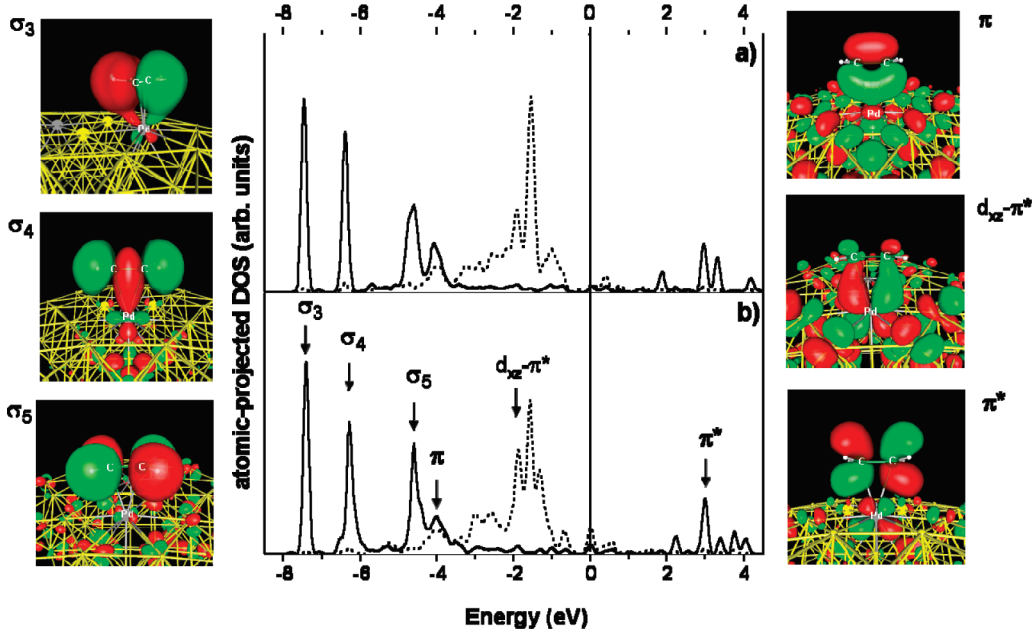


Figure 2. Atom-projected DOS of ethylene molecule (black lines) and Pd d states (dotted black lines) in the C_2H_4/Pd_2Au_{55} (a) and C_2H_4/Pd_2Au_{62} (b) systems. The insets display the full orbital character (contour factor 0.03) of the σ_3, σ_4, and σ_5 peaks (left) and π, π^*, and $d_{\pi z}-\pi^*$ bands (right) of the C_2H_4/Pd_2Au_{62} cluster model. The zero of energy indicates the highest occupied molecular orbital.

significantly decreases by gold addition due to the ligand effect,[33,34] the interaction of the π-bonded ethylene with Pd monomers finally is in the energy regime of the chemisorption/physisorption limit.

Unsurprisingly, this effect has been also observed for the adsorption of C_2H_4 on the PdAu(111) surface. The BE value for this adsorption, calculated with the cluster model used in this study, is 5.5 kcal/mol and is just a little lower than the calculated BEs for the Pd(111) surface, which are in the range of 6.4−7.2 kcal/mol.[45−47] These results are in line with the observations made in the TPD experiments on PdAu(111) surface alloy,[28] which indicated that the heat of adsorption of ethylene decreases monotonically as the gold coverage increases. On the other hand, a desorption activation energy of 13.6−14.8 kcal/mol has been estimated experimentally[28] for the ethylene adsorption on PdAu(111) with medium−high gold coverage (between ∼0.5 and ∼0.7). The computed BE is, therefore, underestimated by about 7−9 kcal/mol with respect to the experimental value. However, it is worth underlining that this discrepancy, beside the DFT limits in the description of such long-range interactions, can be due to the uncertainty associated with the experimental estimation of the heat of adsorption. Indeed, peak broadness, in the TPD spectra, is observed for the ethylene adsorption on PdAu(111) surface at gold coverage higher than 0.73 ML. Furthermore, the stronger interaction between ethylene and the PdAu(100) surface with respect to the PdAu(111) surface is in line with the behavior observed for the adsorption of other organic molecules (such as CO[45,48] and other adsorbed species studied here) on these two surfaces alloys.

The different adsorption orientations for the π-bonded ethylene on the PdAu(111) surface (i.e., with the carbon−carbon axis along the $0\bar{1}0$, $\bar{1}01$, and $\bar{2}11$directions) explored in this study have practically shown the same energetics, proving to be adsorption degenerate states.

The main infrared features of π-bonded ethylene have been detected in the range between 900 and 960 cm^{-1} by both RAIRS[28] and HREELS[25,29] experiments and assigned to the CH_2 wagging modes. In the same spectral region we found the most intense infrared vibrations that just correspond to the in-phase wagging vibrational mode, where in phase indicates the simultaneous wagging of the two CH_2 groups out of the molecular plane. In particular, in the case of the PdAu(111) surface alloys the main infrared feature (with the highest peak intensity, see Table 1) is calculated to lie at 938 cm^{-1}. This value is in excellent agreement with the main experimental feature detected at 931 cm^{-1} by RAIRS[28] for the π-bonded ethylene adsorbed on a PdAu(111) surface with a 0.73 ML gold coverage and annealed to 120 K (to prevent the interference of ethylene molecules adsorbed on Au atoms).

Further remarks regarding the analysis of the π-bonded ethylene vibrational spectra will be given in the next paragraph for a comparison with the calculated vibrational properties of the surface vinyl species.

The fundamental features of ethylene π-bonded adsorption on the Pd noncontiguous pairs can be described in the framework of the classic Dewar−Chatt−Duncanson model of orbital interactions.[49,50] The filled bonding π orbital of ethylene transfers charge to the empty metal d states at the surface, while the occupied metal d band back-donates electron density into the ethylene antibonding π^* orbital. This model is confirmed by our calculations, in agreement with the periodic slab supercell-projected DOS,[33] as shown by the atom-projected DOS for the C_2H_4/Pd_2Au_{55} and the C_2H_4/Pd_2Au_{62} cluster models depicted in Figure 2a and 2b.

In the noninteracting system (where the ethylene molecule is located 4 Å above the substrate) the highest occupied molecular orbital (HOMO) of ethylene and the lowest unoccupied molecular orbital (LUMO) lie at −1.2 and 3.9 eV, respectively. In the equilibrium structure of adsorbed ethylene on PdAu clusters, a relevant depletion of the π-orbital peak of ethylene (HOMO) can be observed at −4.0 eV along with a broadening of the π^* peak (LUMO) into a wide energy range, in accordance with the common donation and back-donation picture (see Figure 2). In general, small bonding and antibonding overlaps between Pd d bands and σ and π states of ethylene have been detected, in line with a Dewar−Chatt−Duncanson description of a weak chemical interaction. The main contribution to the adsorption bond comes from the coupling of the HOMO of ethylene with the empty d_{z^2} orbital of the Pd monomer.

As confirmed by the orbital plot at −2.0 eV ($d_{xz}-\pi^*$), some orbital states with π^* character appear to be occupied, but the feeble overlapping with the metal d states indicates a back-donation mechanism with a limited charge transfer.

According to this description of the Pd−ethylene interaction, a very small elongation of the C−C bond by 0.05 and 0.04 Å with respect to the gas-phase ethylene (1.333 Å) has been observed (see Table 1) on the PdAu(100) and PdAu(111) surfaces, respectively. The calculated Pd−C distances (about 2.3−2.4 Å) and the planar and tilt angles (θ and ϕ, respectively), whose values confirm a negligible distortion of the sp^2 hybridization of the C atoms, confirm the weak chemical interaction between ethylene and Pd atoms in the gold-rich PdAu surface alloys.

3.2. Adsorption of Vinyl. According to the Moiseev-type mechanism[21−23] the vinyl acetate formation involves a surface vinyl species that couples with the coadsorbed acetate. This radical species is a surface intermediate of the dehydrogenation reaction of an adsorbed ethylene molecule. However, it has been generally observed that ethylene decomposition leads to the formation of the stable surface intermediate ethylidyne on the close-packed fcc (111) and hcp (0001) metal surfaces,[51,52] while adsorbed vinylic (vinyl or vinylidene) species seem to be more favorable intermediates only on the more open fcc (100) surfaces.[51−54] On the other hand, the density functional study of the ethylene dehydrogenation pathways over Pd(111)[55] has strongly suggested that the surface vinyl species is even formed on this closed packed surface, but it is unlikely to be detected spectroscopically because it quickly reacts along other surface reaction paths (e.g., rehydrogenation of ethylene or ethylidyne formation[56]). These theoretical results are corroborated by low-energy electron diffraction (LEED) analysis of ethylene decomposition over Pd(111) at low temperature,[57] which suggests the presence of ∼15% of tilted species in the surface layer having the same geometry of a di-σ ($\eta^1\eta^2$) vinyl surface species.

The surface vinyl species formation from adsorbed ethylene, instead, seems to be inhibited over PdAu(001) surface alloys due to the high activation barrier of this elementary step, as recently suggested by slab supercell calculations.[33] It is worthwhile underlining that surface oxygen and hydroxyl groups should be present in the reaction conditions of VAM synthesis[25] leading to water formation. Since surface oxygen atoms promote the β-hydrogen elimination from ethyl acetate-like species (in the Samanos-type mechanism), lowering the activation barrier,[34] these surface species should also facilitate the formation of surface vinyl species by hydrogen abstraction from ethylene (in the Moiseev-type mechanism). Therefore, the theoretical predictions on the adsorption properties of vinyl radical on PdAu surface alloys are not meaningless and have been carefully carried out in this study in order to be possibly compared with future experimental studies.

On the clean Pd(111) surface the vinyl species prefers to adsorb with an $\eta^1\eta^2$(C,C) fashion, showing an approximate sp^3 hybridization at each C atom with a BE of 56−61 kcal/mol, whereas the η^1-atop adsorption mode (sp^2) is less favorable by about 10 kcal/mol, as indicated by both periodic slab and cluster model calculations.[55] All attempts to find local minima related to the $\eta^1\eta^2$ adsorption mode of vinyl on the cluster models of PdAu(100) and PdAu(111) surfaces have been unsuccessful, indicating that the absence of contiguous Pd monomers determines the destabilization of such an $\eta^1\eta^2$ (C,C) adsorption.

The most stable adsorption mode of vinyl on second-neighbor Pd monomer pairs is the σ-bonded η^1-vinyl on top of Pd monomer (denoted as η^1-Pd) with the vinyl molecule vertically adsorbed. The full geometry optimizations starting from the η^1-bridge and η^1-hollow adsorption modes relax to the η^1-atop one. Besides, a stable η^1-vinyl species adsorbed on top of the Au atom (denoted as η^1-Au) has been obtained on both the PdAu(100) and PdAu(111) surfaces having an adsorption energy 10−11 kcal/mol lower than the η^1-Pd vinyl species.

The vinyl radical adsorption properties on the Pd monomers and the Au atoms of both Pd_2Au_{55} and Pd_2Au_{62} cluster models in the stable η^1-atop configurations are listed in Table 2.

For the η^1-atop adsorption mode, all the orientations of the C−C axis with respect to the surface directions have shown the same adsorption energy. The BEs for the η^1-Pd vinyl on PdAu(100) and PdAu(111) surfaces alloys are 44.1 and 41.4 kcal/mol, respectively. These values are slightly lower than the calculated value of 49.5 kcal/mol for the same adsorption mode on the Pd(111) clean surface.[55] Therefore, even if the ligand effect of gold weakens the interaction between the Pd atom and the adsorbed vinyl, this radical species remains strongly bonded to the surface alloys. The C−C bond distance in the adsorbed vinyl species is, in fact, elongated with respect to the gas-phase vinyl molecule (1.309 Å), becoming similar to that of the gas-phase ethylene (1.333 Å). The sp^2 hybridization of both the C atoms in the surface vinyl species is also revealed by the θ(PdCC) angle, whose values (around 127°) are close to the 120° of ethylene. The small values of the tilting angle ϕ_{tilt} (between 2° and 7°) confirm the planarity of the Pd−vinyl system, which always is perpendicular to the surface plane.

Due to the loss of symmetry by the presence of the Pd−C bond instead of the H−C bond and to the different adsorption modes of the ethylene and vinyl surface species, the calculated vibrational spectra for the η^1-vinyl adsorbed

Table 2. BSSE-Corrected Binding Energies (BE, in kcal/mol), Equilibrium Bond Distances (*d*, in Angstroms), Angles (θ and ϕ, in degrees), and Scaled Vibrational Frequencies (in cm^{-1}) of the η^1-Pd and η^1-Au Vinyl Surface Species on the Pd_2Au_{55}(24,21,12) and Pd_2Au_{62}(30,21,13) Cluster Models[a]

	PdAu(100)		PdAu(111)	
	η^1-Pd	η^1-Au	η^1-Pd	η^1-Au
BE	44.1	34.6	41.4	30.3
d(C−C)	1.336	1.331	1.333	1.338
d(S−C)	2.007	2.071	2.016	2.113
θ(SCC)	127.7	125.7	127.5	125.5
ϕ_{tilt}	6.9	2.3	1.6	7.8
$\nu(CH_2)$asymm + ν(CH)	3126(w)	3149(w)	3124(w)	3128(w)
$\nu(CH_2)$symm + $\nu(CH)_{in}$	3099(w)	3119(w)	3090(w)	3114(w)
$\nu(CH_2)$symm + $\nu(CH)_{out}$	3026(w)	3040(w)	3027(w)	3030(w)
ν(CC)	1561(s)	1580(s)	1560(s)	1551(s)
(CH_2) sciss	1356(m)	1366(m)	1352(m)	1361(m)
CH_2CH $rock_{in}$	1154(vs)	1174(vs)	1145(vs)	1177(vs)
CH_2CH $rock_{out}$	929(w)	948(w)	909(w)	954(w)
CH_2CH wag_{in}	905(m)	938(m)	901(m)	952(m)
CH_2CH wag_{out}	842(w)	877(w)	841(w)	895(w)
ν(S−C)	443(w)	475(w)	449(w)	437(w)

[a] The ϕ_{tilt} angle indicates the tilting angle between the vinyl molecular plane and the plane normal to the surface. S indicates the surface atom involved in the adsorption bond (Pd for η^1-Pd and Au for η^1-Au). ν: stretching. Subscript in and out indicate the in-phase and out-of-phase motions of the CH_2 and CH groups. (vs) very strong; (s) strong; (m) medium; (w) weak.

molecule are, obviously, different with respect to those calculated for the π-bonded adsorbed ethylene (compare Table 1 with Table 2). In particular, the main infrared feature for the η^1-Pd vinyl is calculated at 1154 and 1145 cm^{-1} on PdAu(100) and PdAu(111), respectively, and is associated with the in-phase rocking vibration of CH_2 and CH groups, whereas, as already discussed, the main peak of the π-bonded ethylene on the two surface alloys corresponds to the in-phase wagging of CH_2 groups and lies at 922 and 938 cm^{-1}, respectively.

For both PdAu surface alloys, a significant red shift of the vinyl vibration features with respect to the ethylene-adsorbed case has been observed in all regions of the calculated spectra. The C−C stretching and in-plane CH_2 bending modes are largely coupled in a wide-frequency range (1203−1607 cm^{-1}) for the π ethylene, whereas they are red shifted in the range 929−1561 cm^{-1} and become less coupled in the vibrational spectra of the adsorbed vinyl. Furthermore, the C−C stretching mode at about 1560 cm^{-1} appears to be particularly intense in the calculated vibrational spectra of surface vinyl species and should be clearly visible experimentally. The stretching modes of C−H bonds and the out-of-plane bending modes are generally red shifted by 50−100 cm^{-1}. The former modes move from 3080−3184 cm^{-1} in the π-bonded ethylene to 3026−3126 cm^{-1} in the η^1-Pd vinyl, whereas the latter ones move from 841−905 to 811−1005 cm^{-1}. The most consistent red shift has been observed for the Pd−C stretching and moves from 158−175 cm^{-1} for the weak π interaction of ethylene with Pd monomer to 443−449 cm^{-1} for the strong σ bond involved in the Pd−vinyl interaction.

Very similar adsorption geometries have been obtained for the η^1-vinyl adsorbed on Au atoms, as it is shown in Table 2. The main difference in the calculated vibrational spectra concerns a blue shift by 20−30 cm^{-1}, which should be experimentally detectable in the case of the most intense vibrational frequency.

3.3. Adsorption of Acetic Acid. Experimental studies have shown that acetic acid adsorbs molecularly on Pd(111) surfaces and forms η^2-acetate species on heating, even if the formation of η^1-acetate species has not been excluded.[58] In this section we focus on the adsorption of acetic acid molecules on PdAu surface alloys by comparing theoretical results and experimental data.

Indications on the adsorption features of intact acetic acid molecules on PdAu(111) come from RAIRS experiments combined with TPD data.[31] One of the main conclusions of the combined TPD and RAIRS experiments, carried out as a function of gold content, is that acetic acid adsorbs molecularly and desorbs intact for gold mole fractions greater than ∼0.5. TPD spectra collected as a function of gold coverage show that the presence of gold does not change significantly the desorption activation energy (at about 200 K) of the acetic acid intact molecules. Indeed, the acetic acid desorption from the monolayer on Pd-rich PdAu(111) surface[31] and clean Pd(111)[58] and Pd(100)[59] surfaces occurs at about 200 K corresponding to a desorption activation energy of about 12 kcal/mol.

The BE for a di-σ-(O,O)-bonded acetic acid molecule on bridge sites of Pd(111) surface, calculated by use of both cluster and periodic slab supercell models, is in the range of 4.8−6.0 kcal/mol[60] and underestimated with respect to the experimental desorption activation energy.[58] On both Pd and PdAu the predominant form of surface acetic acid species is surface catemers, and the nature of the catemer is gold coverage dependent.[31,58] Since in the theoretical models, both here and in previous studies on Pd surfaces,[60] only the monomeric acetic acid species have been considered, a perfect agreement between the theoretical and experimental results is unlikely.

The acetic acid molecule can adsorb on Pd surfaces in both monodentate and bidentate ways, but periodic slab model calculations[60] have indicated that the bidentate η^2-(O,O) surface species is more stable with respect to the monodentate η^1-(O_C) one (where O_C indicates the carbonyl oxygen atom involved in the Pd−O bond) on the Pd(111) surface.

In our study all possible adsorption geometries of acetic acid on second-neighbor Pd monomer pairs have been considered (see Figure 3). In particular, both the η^2-(O,O) atop (Pd-chelating) and the μ_2-η^2-(O,O) bridge (Pd−Au, not chelating) bidentate adsorption modes have been examined. All attempts to find stable adsorption states having these geometries have been unsuccessful, indicating that bidentate adsorption of acetic acid monomer on isolated Pd sites is really improbable. Moreover, the possibility to find a μ_2-η^2-(O,O)-CH_3COOH bonded to both second-neighbor Pd monomers can be excluded due to the too long Pd−Pd distance in these ensembles (at least 4.08 Å) with respect to the O−O distance in the acetic acid molecule (2.270 Å). All the η^2-CH_3COOH adsorption modes force the adsorbate to obtain its high-energy isomeric configuration (the cis conformer).

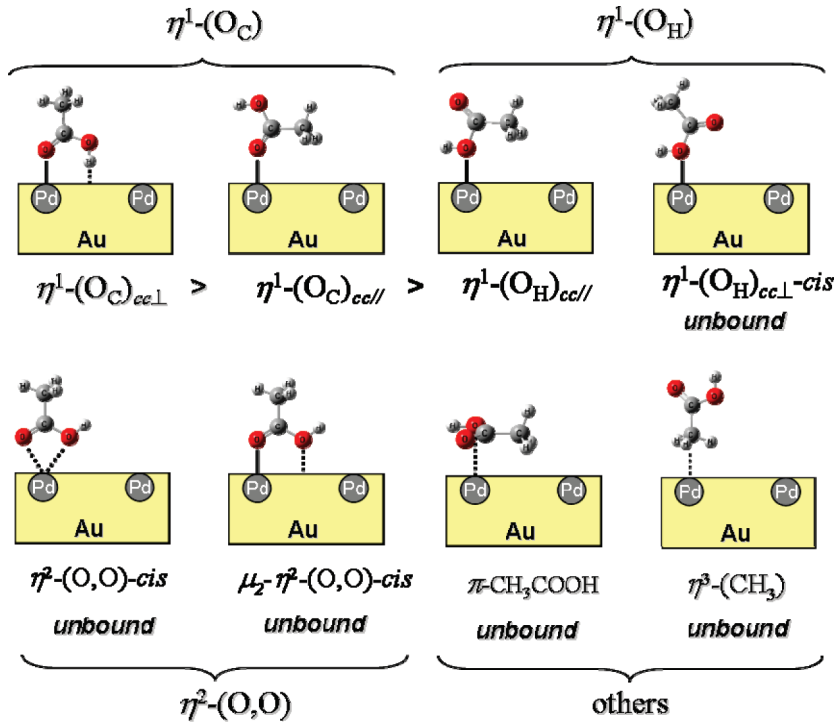


Figure 3. Adsorption geometries of acetic acid on second-neighbor Pd monomer pairs considered in this study. The stability trend on the PdAu surface alloys has been indicated. O_H and O_C specify the oxygen atom involved in the Pd−O bond in the monodentate adsorption, hydroxyl oxygen atom, and the carbonyl oxygen, respectively. The label *cis* indicates the less stable acetic acid conformer; *trans* is omitted.

Table 3. BSSE-Corrected Binding Energies (BE, in kcal/mol), Equilibrium Bond Distances (d, in Angstroms), Angles (θ, in degrees), and Scaled Vibrational Frequencies (in cm^{-1}) of the η^1 Adsorption Modes of Acetic Acid on the Pd_2Au_{55}(24,21,12) and Pd_2Au_{62}(30,21,13) Cluster Models[a]

			PdAu(100)			PdAu(111)		
	gas-phase CH_3COOH	exp[b]	η^1-$(O_C)_{cc\perp}$	η^1-$(O_C)_{cdl}$	η^1-$(O_H)_{cdl}$	η^1-$(O_C)_{cc\perp}$	η^1-$(O_C)_{cdl}$	η^1-$(O_H)_{cdl}$
BE		~12.0	9.0	5.1	2.9	5.6	3.0	2.1
d(Pd−O_C)			2.330	2.374		2.451	2.475	
d(Pd−O_H)					2.549			2.806
d(O−H)	0.980		1.011	0.991	0.988	1.010	0.990	0.990
d(H−S)			1.930		2.510	2.090		2.691
d(C−C)	1.513		1.508	1.501	1.504	1.510	1.504	1.509
d(C−O_C)	1.215		1.239	1.232	1.210	1.233	1.229	1.213
d(C−O_H)	1.365		1.332	1.346	1.388	1.338	1.350	1.376
θ(COH)	105.6		109.6	106.9	106.7	109.1	106.8	107.6
ν(O−H)	3622(m)		3243(vs)	3607(m)	3612(m)	3250(vs)	3611(m)	3607(m)
ν(C=O)	1820(s)	1720 (s)	1677(vs)	1714(vs)	1829(s)	1697(vs)	1727(vs)	1817(s)
$\delta_s(CH_3)$	1422(m), 1415(m)	1426 (br)	1420(w), 1410(m)	1435(m), 1415(m)	1412(m), 1402(m)	1422(w), 1411(m)	1431(m), 1413(m)	1422(m), 1414(m)
ν(C−O)	1371(m), 1309(m)	1294 (br)	1389(s), 1327(w)	1359(s), 1328(m)	1356(m), 1267(m)	1377(s), 1322(w)	1359(s), 1321(m)	1361(m), 1278(m)
γ(OH)	1184(s)	923/958 (s)	1201(m)	1199(s)	1151(s)	1199(m)	1197(s)	1148(s)
$\rho(CH_3)$	1035(m)	1052 (m)	1031(w)	1032(w)	1027(w)	1030(w)	1031(w)	1032(w)

[a] Geometrical parameters and vibrational frequencies for the free acetic acid molecule and available experimental data on PdAu(111) are reported for comparison. [b] Reference 31. ν, stretching; δ, deformation; γ, bending; ρ, rocking modes. For detailed assignments, see ref 61. (vs) very strong; (s) strong; (m) medium; (w) weak; (br) broad.

On the other hand, three monodentate adsorption modes are local minima having different stability on second-neighbor Pd monomer pairs. The adsorption properties of these monodentate acetic acid molecules are listed in Table 3.

Other explored adsorption geometries, such as those with the acetic acid molecular plane parallel to the surface or the methyl group forming a surface bond, are very unstable on the surface alloys studied here (Figure 3).

The monodentate CH_3COOH can bond to the Pd monomers through the hydroxyl oxygen atom (O_H) or the carbonyl oxygen (O_C). On both PdAu(100) and PdAu(111), the η^1-(O_C)-CH_3COOH geometry is more stable than the η^1-(O_H)-CH_3COOH (see Figure 3).

In particular, the η^1-(O_C)-CH_3COOH having the C−C axis perpendicular to the surface plane (denoted as η^1-$(O_C)_{cc\perp}$) is the most stable adsorption state and shows BE values of 5.6

and 9.0 kcal/mol on the PdAu(111) and (100) surfaces, respectively.

In this adsorption geometry the O−H bond points to the hollow site of the surface, forming a very weak hydrogen bond with the gold surface that to some extent stabilizes this adsorption state. The O−H bond is, in fact, slightly elongated to 1.010 Å with respect to the 0.980 Å of the free acetic acid molecule.

The hydrogen−surface distances are about 2.0 Å on both surface alloys, indicating that this adsorption state should also facilitate hydrogen abstraction from acetic acid, leading to the formation of surface acetate species.

The contributions to the adsorbate−surface bonds in the η^1-$(O_C)_{cc\|}$ and η^1-$(O_H)_{cc\|}$ surface species are the weak interactions between the Pd surface atom and the sp^2 and sp^3 oxygen lone pairs of carbonyl and hydroxyl oxygen atoms, respectively. Indeed, on both PdAu surface alloys the binding energies of η^1-$(O_C)_{cc\|}$ and η^1-$(O_H)_{cc\|}$ adsorption states are 2−3 kcal/mol lower than the BEs of η^1-$(O_C)_{cc\perp}$, suggesting that the hydrogen−surface bond in the latter geometry has this strength. The η^1-$(O_H)_{cc\perp}$-CH_3COOH is not stable on the surface alloys because in this adsorption geometry the acetic acid molecule is forced in its *cis* isomeric configuration.

Let us compare now the calculated vibrational spectra of the stable η^1-CH_3COOH adsorption states with the RAIRS spectra collected from the adsorption of acetic acid on the Au-rich PdAu(111) surface alloys[31] and with the gas-phase free molecule vibrations, see Table 3.

First, in the experimental RAIRS spectra[31] two main features appear at 1720 and 923 cm^{-1} corresponding to the carbonyl C−O stretching, denoted as ν(C═O), and the hydroxyl group bending motions, denoted as γ(OH), respectively. The calculated ν(C═O) frequencies for both the η^1-$(O_C)_{cc\|}$ and the η^1-$(O_H)_{cc\perp}$ adsorption states show strong intensities at 1697−1727 cm^{-1}, in accordance with the observed experimental features. On the contrary, for η^1-(O_H)-CH_3COOH the carbonyl C−O stretching is almost 100 cm^{-1} blue shifted (1817 cm^{-1}), becoming very similar to the ν(C═O) in the free acetic acid molecule (1820 cm^{-1}).

On the other hand, the calculated γ(OH) bending frequencies of all the stable η^1-CH_3COOH adsorption states are close to those of gas-phase acetic acid (at 1184 cm^{-1}) and therefore far away from the experimentally observed features between 923 and 958 cm^{-1}. This result comes as no surprise since the RAIRS peaks in this spectral region are exclusively due to the presence of catemeric and dimeric acetic acid species that form on PdAu(111) surface alloy at the considered acetic acid coverage (i.e., 1.2 L). Although it is not the aim of the present paper to follow the variation of the vibrational frequencies of acetic acid due to formation of dimeric or catemeric species, we carried out a test calculation on the acetic acid dimer on the PdAu(111) surface. The calculated γ(OH) bending frequency for one of the possible acetic acid dimer configurations is 942 cm^{-1}, which falls in the range of the experimental values.

Moreover, the C−O_H stretching and the symmetric deformation of the CH_3 group, denoted as ν (C−O) and $\delta_s(CH_3)$, respectively, appear as broad RAIRS peaks at 1294 and 1426 cm^{-1}, and this is in good agreement with the calculated spectra. For all the adsorption states, in fact, these two vibrational modes have medium intensities and are strongly coupled in a wide frequency range between 1267 and 1435 cm^{-1}.

Finally, the stretching mode of the hydroxyl group, ν(O−H), has not been observed in the RAIRS spectra collected for the PdAu(111) surface alloy, while in the HREELS spectra of acetic acid adsorbed on Pd(111) surface[62] a broad peak at 2525 cm^{-1} has been assigned to this stretching mode. The large frequency shift of this mode with respect to the gas-phase acetic acid molecule (calculated at 3622 cm^{-1}) has been suggested to be indicative of the formation of catemers on the clean Pd surface. In the calculated vibrational spectra, as expected, this large frequency shift is not observed. The ν(O−H) vibrations lie at about 3610 cm^{-1} for the η^1-$(O_C)_{cc\|}$ and the η^1-$(O_H)_{cc\|}$ surface species. For the η^1-$(O_H)_{cc\perp}$-CH_3COOH, where a weak hydrogen bond is involved, a significant red shift has been observed in the 3243−3250 cm^{-1} range.

The theoretical cluster model results, therefore, confirm the RAIRS assignments of the ν(C═O), ν(C−O), and $\delta_s(CH_3)$ modes and support the indication of acetic acid catemers formation on PdAu surface alloys because surface acetic acid monomers do not show essential features at 920−960 and ca. 2525 cm^{-1} corresponding to the hydroxyl γ(OH) and ν(O−H) vibrational modes, respectively.

3.4. Adsorption of Acetate. The surface chemistry of acetic acid has been studied on a number of single-crystal surfaces including Pd(111),[58] Pd(110),[63] and clean and oxygen-covered Pd(100)[59] and Pd−Au alloys.[25,29,31] These experiments have indicated that acetate surface species are easily formed, even at room temperature, after acetic acid adsorption on the Pd and PdAu surfaces. Both mono- and bi-dentate spectroscopically detected acetate surface species can decompose to form CO and CO_2 or rehydrogenate to give acetic acid again. The role of Au addition to Pd surfaces is to stabilize the adsorbed acetic acid and acetate surface species by decreasing their decomposition tendency.

Theoretical studies combined with LEED measurements have shown that the preferred adsorption mode of acetate on Pd(111) surface is the bidentate species di-σ-bonded to two Pd atoms on a bridge site having the C−C axis perpendicular to the surface.[64] Binding energies of 50.7 and 52.6 kcal/mol have been calculated by use of cluster and slab supercell models, respectively.[60]

Periodic slab supercell calculations performed on second-neighbor Pd monomer pairs of PdAu(001)[34] surface alloy have shown that the preferred adsorption mode of the acetate is the di-σ-bonded on the PdAu bridge site, denoted as η^2-(O,O), having a BE equal to 47.5 kcal/mol. The π-bonded geometry, with both oxygen atoms bonded to Pd monomer in a bidentate chelating fashion, is less stable by about 13 kcal/mol.

According to these slab supercell calculations, metastable states with chelating geometry that lie approximately 11−12 kcal/mol higher in energy than the corresponding bridge adsorptions have been found on both PdAu(111) and PdAu(100) surfaces by use of cluster models. However, they

Table 4. BSSE-Corrected Binding Energies (BE, in kcal/mol), Equilibrium Bond Distances (*d*, in Angstroms), and Scaled Vibrational Frequencies (in cm^{-1}) of the η^1- and η^2-Acetate on the Pd_2Au_{55}(24,21,12) and Pd_2Au_{62}(30,21,13) Cluster Models[a]

	adsorption mode	BE	d(Pd−O_1)	d(Au−O_2)	d(C−C)	d(C−O_1)	d(C−O_2)
PdAu(100)	η^2-(O,O)	46.4	2.135	2.271	1.523	1.273	1.271
	η^1-(O)	30.5	2.114		1.514	1.321	1.234
slab supercell[b]	η^2-(O,O)	47,5	2.134	2.243			
PdAu(111)	η^2-(O,O)	34.1	2.145	2.356	1.525	1.281	1.265
	η^1-(O)	16.6	2.128		1.512	1.314	1.238

	adsorption mode	ν_{as}(OCO)	ν_s(OCO)	δ_{as}(CH_3)
PdAu(100)	η^2-(O,O)	1482 (m) +24]	1327 (s) [-5]	1417−1437 (m) [+4]
	η^1-(O)	1613 (s)	1210 (w)	1403−1425(m)
PdAu(111)	η^2-(O,O)	1480 (w) [+23]	1315 (s) [−2]	1421−1439 (w) [+4]
	η^1-(O)	1597 (s)	1232 (w)	1410−1434 (m)
PdAu(111) RAIRS[c]	η^2-(O,O)		∼1402(s)	
	η^1-(O)	∼1653 (s)		1426 (s)
PdAu(111) HREELS[d]	η^2-(O,O)	1420(br)		

[a] Differences between negatively charged ($q = -1$) and neutral acetate/cluster systems are reported in brackets. (s) strong; (m) medium; (w) weak; (br) broad. ν_{as} asymmetric stretching; ν_s symmetric stretching; δ deformation; γ bending; ρ rocking modes. [b] Reference 34. [c] Reference 31. [d] References 25 and 29.

do not result as stable local minima because they easily relax to the more stable PdAu-bridge site geometries.

Since the η^1-acetate, which can be spectroscopically distinguished from the bidentate η^2-species, has been experimentally observed to be a stable surface species, the monodentate adsorption mode, denoted as η^1-(O), has also been considered in this study. The adsorption properties, including binding energies, geometrical parameters, and vibrational frequencies, of both bidentate and monodentate modes are listed in Table 4.

It is necessary to underline that the acetate molecule has two stable rotamers having one of the HCC planes perpendicular (staggered) or parallel (eclipsed) to the OCO plane. As already shown for the adsorption of acetate on Au[65] and Ag[66] surfaces, the internal rotation of methyl group has a marginal influence on the calculated vibrational spectra. Therefore, only the more stable staggered isomer has been considered in this study.

η^2-(O,O) is strongly bonded to the palladium−gold surface alloys with binding energies of 46.4 and 34.1 kcal/mol for PdAu(100) and PdAu(111), respectively. The former BE value is in very good agreement with the previous calculated BE on PdAu(100) by the slab supercell approach,[34] while the lower BE value for the more dense PdAu(111) surface is in line with the results obtained for the other adsorbates studied here. According to the decrease of the binding energy, the Pd−O and Au−O distances increase from 2.135 to 2.145 Å and 2.271 to 2.356 Å for the PdAu(100) and the PdAu(111) surfaces, respectively.

The BE differences between the η^2-(O,O) bridge and the η^1-(O) on top adsorptions are around 16−18 kcal/mol, indicating that monodentate acetate is significantly less stable than the bidentate one on PdAu surface alloys. The only stable monodentate adsorption state has the C−C axis parallel to the surface plane.

Despite this trend in BE, shorter Pd−O bond lengths have been observed in the η^1-(O)-acetate (2.114 and 2.128 Å for the PdAu(100) and PdAu(111), respectively) with respect to the η^2-(O,O)-acetate (2.135 and 2.145 Å, respectively). These results suggest that the Au−O bond provides an important contribution to the chemisorption bond of the bidentate surface species. This argument is important since the bidentate adsorbed acetate is a key surface species in the VAM synthesis on PdAu surfaces.[25] The presence of the Au−O bond in the adsorbed surface acetate can offer further (or different) explanations for the promotional role of gold in these Pd-based catalysts and for the different reactivity observed on PdAu(100) and PdAu(111) surfaces. Indeed, the breaking of the Au−O bond instead of the Pd−O bond in the ethylene−acetate coupling reaction on PdAu(001) has been indicated as the primary motivation of the higher reactivity of second-neighbor Pd pairs with respect to contiguous Pd ensembles.[34]

Let us consider now the calculated vibrational spectra of the bidentate and monodentate acetate surface species that can be compared with previous RAIRS[31] and HREELS[25,29] experiments. The major infrared features for the surface acetate species that generally allow distinguishing between bidentate and monodentate species are the symmetric and asymmetric stretching frequencies of the carboxylic group, namely, ν_s(OCO) and ν_{as}(OCO), and the asymmetric deformations of the CH_3 group, i.e., δ_{as}(CH_3).

In the RAIRS experiments the bidentate acetate has been recognized following the ν_s(OCO) symmetric stretching feature at about 1400 cm^{-1}, whereas in the HREELS spectra a broader peak centered at 1420 cm^{-1}, assigned to the coupled (OCO) stretching frequencies and CH_3 deformation modes, has been observed. This difference can be ascribed to the compositions of the surface alloys in these experiments, namely, Pd-rich surfaces in RAIRS (the η^2-acetate is detected only at a Au mole fraction < 0.33)[31] and isolated Pd sites in HREELS.[25,29]

The calculated ν_s(OCO) stretching frequencies on PdAu surfaces are shifted to lower frequencies, at 1315−1327 cm^{-1}, with respect to the experimental features for bidentate acetate, at about 1400 cm^{-1}. On both clean Pd surfaces[58] and Pd-rich PdAu alloys[31] as well as on clean Au surfaces[65] the observed and calculated ν_s(OCO) stretching frequencies

lie at about 1400 cm^{-1}. The difference between the asymmetric di-σ bonds on the Pd−Au bridge site (see Pd−O and Au−O distances in Table 4) and the symmetrical di-σ bonds on the homonuclear Pd−Pd and Au−Au bridge sites can explain this discrepancy. In fact, the calculated ν_s(OCO) at 1315−1327 cm^{-1}, the $\delta_{as}(CH_3)$ in the 1417−1439 cm^{-1} frequency range, and the ν_{as}(OCO) at about 1480 cm^{-1} (see Table 4) are in good agreement with the broad HREELS band of the "ν (OCO) + $\delta(CH_3)$" modes, centered at 1420 cm^{-1}, observed for the Pd isolated sites.[25,29]

The formation of monodentate acetate on PdAu(111) has been recognized by RAIRS experiments[31] by means of two features at 1653 and 1426 cm^{-1} assigned to the asymmetric ν_{as}(OCO) stretching and the asymmetric methyl deformation $\delta_{as}(CH_3)$, respectively.

In accordance with Tysoe et al.[31] we found that in the monodentate adsorption mode the C−C axis is parallel to the surface plane. Indeed, the asymmetric CH_3 deformation modes lie at 1403−1434 cm^{-1} (about 1426 cm^{-1} in the RAIRS experiments), whereas the symmetric $\delta_s(CH_3)$ deformations are very weak and lie at about 1350 cm^{-1}.

The ν_{as}(OCO) stretching frequencies in the η^1-acetate become strong in intensity and blue shifted to 1597−1613 cm^{-1} with respect to those of the bidentate acetate.

For the most stable η^2-(O,O) adsorption states on PdAu(100) and PdAu(111) surfaces, both the radical and anionic acetate species have been considered in order to evaluate the effect of surface acetate species discharge on the vibrational properties. As already obtained for acetate adsorption on gold[65] and silver[66] surfaces the frequency shifts due to the cluster charge variation are small for the ν_s(OCO) stretching frequencies (∼20 cm^{-1}) and negligible for the $\delta_{as}(CH_3)$ modes (2−5 cm^{-1}).

4. Conclusions

In this work we studied the adsorption properties of ethylene, vinyl, acetic acid, and acetate species on the PdAu(111) and PdAu(100) surface alloys by the density functional cluster model approach. In particular, cluster models have been built up to represent second-neighbor Pd pairs on both surfaces since these are the critical Pd ensembles in the vinyl acetate monomer synthesis, which involves the surface species under study.

In accordance with experimental evidence and previous slab supercell calculations, our cluster model study has shown that ethylene is π bonded on Pd monomers, whereas acetate adsorbs in a bidentate fashion on a palladium−gold bridge site.

The presence of a gold−oxygen bond in the adsorbed surface acetate is confirmed by our cluster model results and should stimulate further rationalizations of the promotional role of gold in the PdAu-catalyzed VAM synthesis.

The calculated binding energies for ethylene and acetate are in very good agreement with preceding slab supercell calculations on the PdAu(100) surface, and the calculated vibrational spectra confirm the experimental frequency assignments, giving more details regarding all the involved vibrational modes.

Even if the surface vinyl species has not been detected on PdAu surface alloys, theoretical predictions on the adsorption properties of the vinyl radical on second-neighbor Pd pairs have been reported in order to be possibly compared with future experimental studies. The surface vinyl species has shown, in fact, strong interactions with Pd monomers and also with Au surface atoms. The calculated vibrational spectra are quite different from those of the adsorbed ethylene molecules.

Density functional cluster model results on the adsorption of acetic acid monomers have confirmed the experimental assignments of the major vibrational modes in the infrared spectra, supporting the indication that acetic acid catemers form on PdAu surface alloys. However, at very low acetic acid coverages, monomeric species should be present on PdAu surfaces, and our results suggest that it adsorbs in a monodentate fashion, forming a weak hydrogen bond with the surface.

In general, the adsorption of all molecules on the bimetallic PdAu surfaces is a less efficient process with respect to the clean Pd surfaces. The ligand effect of the gold atoms surrounding the Pd monomers lowers the adsorption energies. This effect is more pronounced for the more dense PdAu(111) surface alloy.

Finally, the adopted cluster models have been demonstrated to be sufficiently reliable by giving accurate results on the adsorption properties, including binding energies, preferred adsorption sites, stable bonding modes, geometrical parameters, and vibrational frequencies, in full agreement with available slab supercell calculations and the experimental evidence.

Acknowledgment. The authors thank the Università della Calabria for financial support.

References

(1) Sinfelt, J. H. *Bimetallic Catalysts: Discoveries, Concepts and Applications*; Wiley: New York, 1983.

(2) Somorjai, G. *Introduction to Surface Chemistry and Catalysis*; John Wiley & Sons: New York, 1994.

(3) Rodriguez, J. A. *Surf. Sci. Rep.* **1996**, *24*, 223–287.

(4) Moss, R. L.; Whally, L., *Advances in Catalysis*; Academic Press: New York, 1972.

(5) Baddeley, C. J.; Tikhov, M.; Hardacre, C.; Lomas, J. R.; Lambert, R. M. *J. Phys. Chem.* **1996**, *100*, 2189–2194.

(6) Baddeley, C. J.; Ormerod, R. M.; Stephenson, A. W.; Lambert, R. M. *J. Phys. Chem.* **1995**, *99*, 5146–5151.

(7) Han, Y. F.; Kumar, D.; Goodman, D. W. *J. Catal.* **2005**, *230*, 353–358.

(8) Enache, D. I.; Edwards, J. K.; Landon, P.; Solsona-Espriu, B.; Carley, A. F.; et al. *Science* **2006**, *311*, 362–365.

(9) Trimm, D. L.; Önsan, Z. I. *Catal. Rev.* **2001**, *43*, 31–84.

(10) Bonarowska, M.; Malinowski, A.; Juszczyk, W.; Karpinski, Z. *Appl. Catal., B* **2001**, *30*, 187–193.

(11) Li, Z.; Gao, F.; Wang, Y.; Calaza, F.; Burkholder, L.; Tysoe, W. T. *Surf. Sci.* **2007**, *601*, 1898–1908.

(12) Yi, C. W.; Luo, K.; Wei, T.; Goodman, D. W. *J. Phys. Chem. B* **2005**, *109*, 18535–18540.

(13) Legawiec-Jarzyna, M.; Srebowata, A.; Karpinski, Z. *React. Kinet. Catal. Lett.* **2003**, *79*, 157–161.

(14) Venezia, A. M.; La Parola, V.; Pawelec, B.; Fierro, J. L. G. *Appl. Catal., A* **2004**, *264*, 43–51.

(15) Maroun, F.; Ozanam, F.; Magnussen, O. M.; Behm, R. J. *Science* **2001**, *293*, 1811–1814.

(16) Sarkany, A.; Horvath, A.; Beck, A. *Appl. Catal., A* **2002**, *229*, 117–125.

(17) Hilaire, L.; Legare, P.; Holl, Y.; Maire, G. *Surf. Sci.* **1981**, *103*, 125–140.

(18) Jablonski, A.; Overbury, S. H.; Somorjai, G. A. *Surf. Sci.* **1977**, *65*, 578–592.

(19) Provine, W. D.; Mills, P.; Lerov, J. J. *Stud. Surf. Sci. Catal.* **1996**, *101*, 191–200.

(20) Samanos, B.; Boutry, P.; Montarnal, R. *J. Catal.* **1971**, *23*, 19–30.

(21) Moiseev, I.; Vargaftic, M. N.; Syrkin, Y. L. *Dokl. Akad. Nauk SSSR* **1960**, *133*, 377–380.

(22) Moiseev, I. *Catalytic Oxidation*; Sheldon, R. A., Ed.; World Scientific, 1995; p 203.

(23) Nakamura, S.; Yasui, T. *J. Catal.* **1970**, *17*, 366–374.

(24) Stacchiola, D.; Calaza, F.; Burkholder, L.; Tysoe, W. T. *J. Am. Chem. Soc.* **2004**, *126*, 15384–15385.

(25) Chen, M.; Kumar, D.; Yi, C. W.; Goodman, D. W. *Science* **2005**, *310*, 291–293, including supporting online material.

(26) Han, P.; Axnanda, S.; Lyubinetsky, I.; Goodman, D. W. *J. Am. Chem. Soc.* **2007**, *129*, 14355–14361.

(27) Luo, K.; Wei, T.; Yi, C. W.; Axnanda, S.; Goodman, D. W. *J. Phys. Chem. B* **2005**, *109*, 23517–23522.

(28) Calaza, F.; Gao, F.; Li, Z.; Tysoe, W. T. *Surf. Sci.* **2007**, *601*, 714–722.

(29) Chen, M. S.; Luo, K.; Wei, Z.; Yan, Z.; Kumar, D.; Yi, C. W.; Goodman, D. W. *Catal. Today* **2006**, *117*, 37–45.

(30) Owens, T. G.; Jones, T. E.; Noakes, T. C. Q.; Bailey, P.; Baddeley, C. J. *J. Phys. Chem. B* **2006**, *110*, 21152–21160.

(31) Calaza, F.; Gao, F.; I, Z; Tysoe, W. T. *Surf. Sci.* **2007**, *601*, 1351–1357.

(32) García-Mota, M.; Lopez, N. *J. Am. Chem. Soc.* **2008**, *130*, 14406–14407.

(33) Yuan, D.; Gong, X.; Wu, R. *Phys. Rev. B* **2007**, *75*, 233401–233404.

(34) Yuan, D.; Gong, X.; Wu, R. *J. Phys. Chem. C* **2008**, *112*, 1539–1543.

(35) Mazzone, G.; Rivalta, I.; Russo, N.; Sicilia, E. *J. Phys. Chem. C* **2008**, *112*, 6073–6081.

(36) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(37) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(38) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(39) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(40) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(41) Andrae, D.; Hæussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123–141.

(42) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(43) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(44) Neugebauer, J.; Hess, B. A. *J. Chem. Phys.* **2003**, *118*, 7215–7225.

(45) Ge, Q.; Neurock, M. *Chem. Phys. Lett.* **2002**, *358*, 377–382.

(46) Neurock, M.; van Santen, R. A. *J. Phys. Chem. B* **2000**, *104*, 11127–11145.

(47) Mei, D.; Hansen, E. W.; Neurock, M. *J. Phys. Chem. B* **2003**, *107*, 798–810.

(48) Yuan, D.; Gong, X.; Wu, R. *Phys. Rev. B* **2007**, *75*, 085428−085432.

(49) Dewar, M. J. S. *Bull. Soc. Chim. Fr.* **1951**, *18*, C71−C79.

(50) Chatt, J.; Duncanson, L. A. *J. Chem. Soc.* **1953**, *75*, 2939–2947.

(51) Masel, R. I. *Principles of Adsorption and Reaction on Solid Surfaces*; John Wiley and Sons: New York, 1996.

(52) Yagasaki, E.; Masel, R. *Catalysis* **1994**, *11*, 165–222.

(53) Stuve, E. M.; Madix, R. J. *J. Phys. Chem.* **1985**, *89*, 105–112.

(54) Zaera, F.; Hall, R. B. *Surf. Sci.* **1987**, *180*, 1–18.

(55) Pallassana, V.; Neurock, M.; Lusvardi, V. S.; Lerou, J. J.; Kragten, D. D.; van Santen, R. A. *J. Phys. Chem. B* **2002**, *106*, 1656–1669.

(56) Azad, S.; Kaltchev, M.; Stacchiola, D.; Wu, G.; Tysoe, W. T. *J. Phys. Chem. B* **2000**, *104*, 3107–3115.

(57) Stacchiola, D.; Calaza, F.; Zheng, T.; Tysoe, W. T. *J. Mol. Catal A: Chem.* **2005**, *228*, 35–45.

(58) Haley, R. D.; Tikhov, M. S.; Lambert, R. M. *Catal. Lett.* **2001**, *76*, 125–130.

(59) Li, Z.; Gao, F.; Tysoe, W. T. *Surf. Sci.* **2008**, *602*, 416–423.

(60) Pallassana, V.; Neurock, M. *J. Catal.* **2002**, *209*, 289–305.

(61) Burneau, A.; Génin, F.; Quilès, F. *Phys. Chem. Chem. Phys.* **2000**, *2*, 5020–5029.

(62) Davis, J. L.; Barteau, M. A. *Langmuir* **1989**, *5*, 1299–1309.

(63) Bowker, M.; Morgan, C.; Couves, J. *Surf. Sci.* **2004**, *555*, 145–156.

(64) James, J.; Saldin, D. K.; Zheng, T.; Tysoe, W. T.; Sholl, D. S. *Catal. Today* **2005**, *105*, 74–77.

(65) Bernà, A.; Delgado, J. M.; Orts, J. M.; Rodes, A.; Feliu, J. M. *Electrochim. Acta* **2008**, *53*, 2309–2321.

(66) Delgado, J. M.; Rodes, A.; Orts, J. M. *J. Phys. Chem. C* **2007**, *111*, 14476–14483.

CT9000137

JCTC Journal of Chemical Theory and Computation

Optimization of Spin-Unrestricted Density Functional Theory for Redox Properties of Rubredoxin Redox Site Analogues

Shuqiang Niu,[†] Jeffrey A. Nichols,[‡,§] and Toshiko Ichiye*,[†]

Department of Chemistry, Georgetown University, Washington, DC 20057-1227, and Environmental Molecular Science Laboratory, Battelle, Pacific Northwest National Laboratory, Richland, Washington 99352

Received August 29, 2008

Abstract: Quantum chemical calculations of metal clusters in proteins for redox studies require both computational feasibility as well as accuracies of at least ~50 mV for redox energies but only ~0.05 Å for bond lengths. Thus, optimization of spin-unrestricted density functional theory (DFT) methods, especially the hybrid generalized gradient approximation functionals, for energies while maintaining good geometries is essential. Here, different DFT functionals with effective core potential (ECP) and full core basis sets for $[Fe(SCH_3)_4]^{2-/1-/0}$ and $[Fe(SCH_3)_3]^{1-/0}$, which are analogues of the iron−sulfur protein rubredoxin, are investigated in comparison to experiment as well as other more computationally intensive electron correlation methods. In particular, redox energies are calibrated against gas-phase photoelectron spectroscopy data so no approximations for the environment are needed. B3LYP gives the best balance of accuracy in energy and geometry as compared to B97gga1 and BHandH and is better for energies than Møller−Plesset perturbation theory series (MP2, MP3, MP4SDQ) and comparable to coupled cluster [CCSD, CCSD(T)] methods. Of the full core basis sets tested, the 6-31G** basis sets give good geometries, and addition of diffuse functions to only the sulfur significantly improves the energies. Moreover, a basis set with an ECP on only the iron gives less accurate but still reasonable geometries and energies.

Introduction

Proteins containing iron−sulfur clusters, which consist of one to four irons tetrahedrally coordinated by cysteine residues and inorganic sulfur atoms,[1] are an important class of electron transfer proteins in a wide variety of processes such as photosynthesis, respiration, nitrogen fixation, and hydrogen metabolism. They also play critical roles in enzymatic processes, regulation of gene expression, generation of radicals, and delivery of sulfur and iron for the synthesis of other proteins.[2−5] Because redox chemistry is important for many of these proteins, identification of the molecular determinants of the reduction potentials of iron−sulfur redox sites is important for a fundamental understanding of their function as well as for engineering reduction potentials by rational design of mutations.

The physical origins of differences in reduction potentials of proteins are often difficult to identify due to the complexity of the protein environment. The reduction potential E° can be obtained from the free energy $\Delta G°$ of the reduction reaction:

$$\text{protein}-[\text{redox site}]^{(n-1)-} \xrightarrow{e^-} \text{protein}-[\text{redox site}]^{n-} \quad (1)$$

A common approach is to divide $\Delta G°$ into the intrinsic reduction free energy (ΔG_{int}) of the redox site independent of the protein, the extrinsic reduction free energy (ΔG_{env}) produced by surrounding protein and solvent at the redox

* Corresponding author e-mail: ti9@georgetown.edu.
[†] Georgetown University.
[‡] Pacific Northwest National Laboratory.
[§] Current address: Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164.

10.1021/ct800357c CCC: $40.75
Published on Web 04/23/2009

site, and the perturbative interaction between the two ($\Delta G_{int/env}$):[6]

$$-nFE^\circ = \Delta G^\circ = \Delta G_{int} + \Delta G_{env} + \Delta G_{int/env} \quad (2)$$

where n is the number of electrons, and F is Faraday's constant. If $\Delta G_{int/env}$ is small, then ΔG_{int} is approximately the reduction free energy of the clusters in the gas phase. Given the complexity of the protein environment, computational methods can play an important role, and it is essential that quantum chemical methods reproduce ΔG_{int} for analogues in the gas phase before attempting to apply them in the protein environment. The key issue is to obtain ΔG_{int} with accuracies of at least ~50 mV. In addition, while geometry is obviously an important indicator of the quality of a calculation, accuracies of only ~0.05 Å in bond lengths are more than sufficient for modeling redox effects on the proteins. Moreover, systematic errors in geometry are preferable to random errors because, for instance, the change in Fe−S bond length upon reduction indicates the increase in ferrous character of the bond.

Significant progress in quantum chemical calculations of metalloprotein redox sites has been made. Density functional theory (DFT)[7] is now recognized as both more efficient and more accurate for treating transition-metal systems than some conventional ab initio molecular orbital (MO) methods such as Hartree−Fock (HF) and Møller−Plesset perturbation (MPn) methods, especially for systems containing first-row transition metals.[8] However, because Mössbauer, electron paramagnetic resonance (EPR), and PES experiments have revealed unusual electronic structures of iron−sulfur clusters with spin-coupled interactions between iron sites,[2−4] spin-unrestricted methods are required. The spin−spin interactions may be treated by the broken-symmetry (BS) DFT approach,[9] in which the ground state is based on a single determinant wave function. Noodleman and co-workers have investigated iron−sulfur clusters constrained to experimental geometries using local density approximation (LDA) DFT with nonlocal (NL) corrections for exchange and BS-DFT and spin projection methods, which give good electronic structures and Heisenberg spin coupling contributions to magnetic properties.[10,11] However, LDA and nonhybrid generalized gradient approximation (GGA) methods usually give good geometries but poor energies,[7,12] and systematic studies of the accuracy of the hybrid GGA functionals in BS-DFT for both structural and energetic properties are lacking despite many DFT studies of iron−sulfur proteins and analogues.[13−15]

Recently, understanding of the intrinsic energetics has been greatly enhanced by gas-phase photoelectron spectroscopy (PES), which is a powerful experimental technique for determining the vertical and adiabatic detachment energies (VDE and ADE, respectively)[16,17] of oxidation. Because the entropic contribution is small, ΔG_{int} is approximately equal to the intrinsic reduction energy ΔE_{int}:

$$\Delta G_{int} \approx \Delta E_{int} = -\mathrm{ADE} = -(\mathrm{VDE} + \lambda_{oxd}) \quad (3)$$

where λ_{oxd} is the intramolecular relaxation energy of the reaction. The calibration of gas-phase calculations against gas-phase PES data rather than electrochemical redox potentials in solution means no approximations are needed for the environment. This type of calibration of DFT calculations for iron−sulfur clusters is leading to several important types of studies. First, DFT calculations can be used to study a wide range of clusters and their redox couples[18−21] because PES experiments are limited to studies of oxidation of analogues that are sufficiently stable in the gas phase, which are not always the relevant redox couples in proteins. Moreover, our calculations indicate that absolute reduction potentials obtained from gas-phase calculations of ΔG_{int} and continuum electrostatic calculations of ΔG_{ext} are in good agreement with experiment.[22] Second, combined DFT/PES studies of analogues with specially designed ligands can be used to study different physical effects individually such as hydrogen bonds on reduction potentials[20] and ligand differences on fission. Finally, the calibration is important for QM/MM simulations of iron−sulfur proteins.

The simplest iron−sulfur redox site is $[Fe(SCys)_4]^{1-/2-}$, which has a single iron core referred to as [1Fe−0S], or [1Fe] for short, and is found in the small proteins rubredoxin (Rd) and desulforedoxin as well as in larger proteins such as desulfoferrodoxin, rubrerythrin, and nigerythrin.[23] Calibration of DFT methods for analogues of this site is important because their BS ground state is a pure high-spin state without any spin contamination. Here, a systematic study is reported of the performance of hybrid GGA methods with different basis sets in comparison to experimental measurements for structural and energetic properties of $[Fe(SCH_3)_4]^{n-}$ and $[Fe(SCH_3)_3]^{n-}$. The energies were calibrated against recent gas-phase PES measurements of the electronic structure, VDE, and ADE by Wang and co-workers.[16,17] In addition, the energies were compared to calculations at the Møller−Plesset perturbation theory series and coupled cluster levels of theory.

Computational Details

The calculations were performed for the $[Fe(SCH_3)_4]^{2-/1-/0}$ and $[Fe(SCH_3)_3]^{1-/0}$ analogues of the rubredoxin redox site. The replacement of the cysteine ligands by thiolate SCH_3 groups has been shown to be a good model for geometries.[9] Here, the D_{2d} ML_4 structure, which is an idealization of the geometry of the Rd redox site,[24] was utilized for the initial structures of $[Fe(SCH_3)_4]^{2-/1-/0}$. In addition, the D_{3h} ML_3 structure, which is an idealization of the experimental geometry of $[Fe(SC_6H_2\text{-}2,4,6\text{-}t\mathrm{Bu}_3)_3]^{1-}$,[25] was utilized for the initial structures of $[Fe(SCH_3)_3]^{1-/0}$.

Three density functionals with better performance in geometry and energetics for small molecules[12] were examined: Becke's original three parameter fit[26] using Lee−Yang−Parr correlation functional (B3LYP),[27] Becke's "half and half" functionals (BHandH),[26] and the modified Becke's 10-parameter functional (B97gga1).[12,28] In addition, seven different basis sets were examined. In the first, referred to here as ECPDZ, the iron is represented by the double-ζ (DZ) effective core potential (ECP) basis set by Couty and Hall[29] modified from the Hay and Wadt basis set.[30] The carbons and sulfurs have the ECPs and double-ζ basis sets with polarization functions (ECPDZP) of Stevens, Basch, and Krauss,[31] and hydrogens have the Dunning−Huzinaga (31) double-ζ basis set.[32,33] In the second, referred to here as

ECPDZ+DZVP, the iron has the same basis set as in ECPDZ, and the basis sets for the hydrogen, carbon, and sulfur atoms are the all-electron Dgauss DZVP polarized DFT orbital basis sets.[34] The third and fourth, DZVP2 and 6-31G**, respectively, are the all-electron Dgauss DZVP2 polarized DFT orbital basis sets[34] and the standard 6-31G basis sets,[35,36] respectively. In the fifth and sixth, referred to here respectively as DZVP2(++)$_S$ and 6-31(++)$_S$G**, sp-type diffuse functions from the 6-31++G** basis sets[35,37] are added to the sulfur atoms of DZVP2 and 6-31G**, respectively. In the seventh, referred to here as 6-31G(3df)$_S$**, three d-type and one f-type polarization functions from the 6-31G(3df,3pd) basis set[35,36] are added to the sulfur atoms of 6-31G**.

The calculated VDE and ADE were refined at the different levels of theory with different basis sets. Because the energies were recalculated for certain geometries, the notation Method(E)/Basis(E)//Method(G)/Basis(G) will be used to denote an energy calculation at the Method(E) level with the Basis(E) sets for the geometry optimized at the Method(G) level with the Basis(G) sets. In addition to the hybrid GGA functionals, the energies were also recalculated at the frozen core MP2, MP3, MP4SDQ, coupled cluster singles and doubles (CCSD), and CCSD with perturbative corrections for triples [CCSD(T)] levels.[38] The zero point energy (ZPE) corrections were taken into account only in the final results because they are small and do not significantly affect the comparisons of calculations reported here.

All quantum mechanics calculations were performed using the NWChem[39] and Gaussian 03[40] program packages.

Results and Discussion

A detailed analysis of the electronic structures of the analogues studied here has been presented elsewhere;[18] the results are summarized briefly here to aid the discussion. In a roughly T_d weak ligand field[41] of the tetrahedral $[Fe(SCH_3)_4]^{n-}$ site, the t_2 and e orbitals have a small splitting energy so that the iron high-spin state is favorable. Taking exchange interaction between α and β spins into account, the spin polarization splits the occupied α-spin orbitals below the β-spin orbitals. Consequently, the molecular orbital interaction between the high-spin iron and the thiolates leads to a less common inverted level pattern, in which the S(3p) ligand orbitals lie below the Fe(3d) minority β-spin orbitals and above the Fe(3d) majority α-spin orbitals.[10] The oxidation from $[Fe(SCH_3)_4]^{2-}$ to $[Fe(SCH_3)_4]^{1-}$ involves a high-lying Fe(3d) minority β-spin electron orbital, while the oxidation from $[Fe(SCH_3)_4]^{1-}$ to $[Fe(SCH_3)_4]^{0}$ involves a ligand orbital with S(3p) α-spin character. Overall, $n = 2$, 1, and 0 oxidation states have high-spin ground states with $S = 2$, 5/2, and 2, respectively, and the major unpaired electron densities localize on the metal orbitals regardless of oxidation state. Thus, both $[Fe(SCH_3)_4]^{1-}$ and $[Fe(SCH_3)_4]^{2-}$ retain tetrahedral character, while $[Fe(SCH_3)_4]^{0}$ becomes a distorted tetrahedral structure (C_{2v}) because of strong Jahn−Teller distortion arising from the oxidation of the degenerate S(3p) orbitals.

Similarly, in the roughly D_{3h} weak ligand field[41] of the triangular $[Fe(SCH_3)_3]^{n-}$ site, the Fe 3d orbitals transform into e″, a_1', and e′, and spin polarization splits the occupied α-spin orbitals below the β-spin orbitals. The $n = 1$ and 0 oxidation states have high-spin ground states with $S = 2$ and 5/2, respectively,[18] and the major unpaired electron densities localize on the metal orbitals in both oxidation states. The oxidation from $[Fe(SCH_3)_3]^{1-}$ to $[Fe(SCH_3)_3]^{0}$ involves a low-lying Fe d_{xz} minority β-spin electron, which has Fe−S π^* antibonding character, so that $[Fe(SCH_3)_3]^{0}$ is planar.

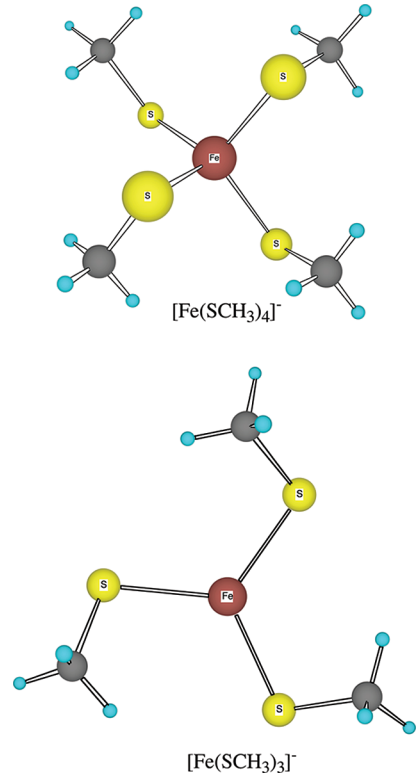


Figure 1. Ball-and stick rendering of $[Fe(SCH_3)_4]^{1-}$ and $[Fe(SCH_3)_3]^{1-}$ in B3LYP/6-31G(3df)$_S$** geometry (Table 1).

Dependence of Molecular Geometries of $[Fe(SCH_3)_4]^{2-/1-/0}$ and $Fe(SCH_3)_3]^{1-/0}$ on Basis Set and Density Functional. The dependence of the geometry on the functional and the basis set was examined and compared to experimental crystal structures. Although PES data are available for $[Fe(SCH_3)_4]^{1-/0}$ and $[Fe(SCH_3)_3]^{1-/0}$,[16,17] experimental structures are not available for the oxidized state of the former or either state of the latter so the experimental structures of $[Fe(S_2\text{-}o\text{-}xyl)_2]^{2-/1-}$,[42] $[Fe(SCH_3)_4]^{1-}$,[43] [Fe(1,2-benzenedithiolate)$_2$(PMe$_3$)]^0,[25] and $[Fe(SC_6H_2\text{-}2,4,6\text{-}t\text{-}Bu_3)_3]^{1-}$[24] were used. However, [Fe(1,2-benzenedithiolate)$_2$(PMe$_3$)]^0 is only moderately representative of $[Fe(SCH_3)_4]^{0}$ because the FeS_4 is nearly planar in the former with a $S = 1$ spin state, whereas it is tetragonal with a high spin ($S = 2$) ground state in the latter. The calculated geometries of $[Fe(SCH_3)_4]^{1-}$ and $[Fe(SCH_3)_3]^{1-}$ are shown in Figure 1 for reference. Overall, the BHandH, B3LYP, and B97gga1 optimized geometries using either ECP or full basis sets of $[Fe(SCH_3)_4]^{2-/1-/0}$ (Table S1−5) and $[Fe(SCH_3)_3]^{1-/0}$ (Tables S1−5, no BHandH) were closer to experiment than previous calculations at the UHF and MP2 levels,[18] although all still slightly overestimated the bond lengths.

Table 1. Geometries for Different Basis in B3LYP Calculations of $[Fe(SCH_3)_4]^{n-}$ and $[Fe(SCH_3)_3]^{n-}$

	basis set	r_{Fe-S}	r_{S-C}	θ_{S-Fe-S}	θ_{Fe-S-C}	$\phi_{C-S-Fe-S}$
$[Fe(SCH_3)_4]^{2-}$	ECPDZ+DZVP	2.440	1.850	110.0	101.9	59.6
	6-31G**	2.414	1.844	110.5	102.8	59.0
	6-31G(3df)$_S$**	2.395	1.836	110.0	102.4	59.4
	exp.[a]	2.356 (±0.032)	1.835 (±0.022)	109.5 (±6.0)	108.2 (±2.4)	NA
$[Fe(SCH_3)_4]^{1-}$	ECPDZ+DZVP	2.335	1.845	110.4	100.8	60.0
	6-31G**	2.314	1.844	110.7	101.7	61.0
	6-31G(3df)$_S$**	2.306	1.833	110.0	101.5	57.2
	exp.[a]	2.267 (±0.015)	1.840 (±0.023)	109.5 (±3.7)	101.5 (±3.7)	NA
	exp.[b]	2.264 (±0.001)	1.753 (±0.008)	107.1 (±0.0)	103.9 (±0.3)	−54.5
$[Fe(SCH_3)_4]^{0}$	ECPDZ+DZVP	2.245	1.839	127.3	105.5	64.8
	6-31G**	2.224	1.841	127.8	107.1	66.7
	6-31G(3df)$_S$**	2.211	1.828	130.3	105.5	66.4
	exp.[c]	2.185 (±0.055)	1.753 (±0.006)	NA	NA	NA
$[Fe(SCH_3)_3]^{1-}$	ECPDZ+DZVP	2.346	1.849	120.5	102.4	3.0
	6-31G**	2.309	1.849	120.5	103.3	31.3
	6-31G(3df)$_S$**	2.295	1.838	120.2	104.0	29.9
	exp.[d]	2.274 (±0.011)	1.796 (±0.009)	120.0 (±3.4)	115.3 (±2.0)	2.0 (±3.0)
$[Fe(SCH_3)_3]^{0}$	ECPDZ+DZVP	2.244	1.844	120.4	106.2	0.5
	6-31G**	2.209	1.846	120.0	107.4	1.4
	6-31G(3df)$_S$**	2.202	1.835	120.5	107.3	0.5

[a] Reference 42. [b] Reference 43. [c] Reference 25. [d] Reference 24.

Focusing first on the functionals, B3LYP and B97gga1 were general superior to BHandH for geometries regardless of the basis set. An important geometric feature that the calculations must reproduce is the Fe−S bond length (Figure 2, Table S1). B97gga1 generally gave the best Fe−S bond lengths, while B3LYP tended to be about 0.02 Å longer. However, the reduced sites, $[Fe(SCH_3)_4]^{2-}$ and $[Fe(SCH_3)_3]^{1-}$, appeared to be the most sensitive to the functional so that B3LYP was better at reproducing the decrease in bond length upon oxidation. The other internal coordinates generally showed less variation with the type of hybrid GGA functional (Tables S2−5). The largest variation was for the thiolate ligand torsion angle, $\phi_{C-S-Fe-S}$, of $[Fe(SCH_3)_3]^{1-}$. Only B3LYP/ECPDZ+DZVP, B97gga1/ECPDZ+DZVP, and B97gga1/6-31G(3df)$_s$** gave values of ∼4°, similar to the experimental value of 2°,[24] whereas the other methods gave values between 30° and 41°, apparently due to stabilization by decreasing the Fe−S π antibonding interaction.[18] On the other hand, the small torsion in the experimental complex, $[Fe(SC_6H_2$-2,4,6-tBu$_3)_3]^{1-}$, appears to caused by steric interference of the ligands so that it is questionable that $[Fe(SCH_3)_3]^{1-}$ would have a small torsion also. In addition, the CCSD(T)/6-31G** energy of the B97gga1/6-31G** geometry was higher than the CCSD(T)/6-31G** energy of the B3LYP/6-31G** structure (Table S7), indicating that CCSD(T) found the B3LYP geometries more favorable. Moreover, B97gga1 is shown below to give poor energetics.

Now focusing on the basis sets, the larger basis sets generally gave superior geometries, as expected. The Fe−S bond lengths (Figure 2, Table S1) tended to improve as the basis set size increased especially on going to the full core basis sets, but again the reduced $[Fe(SCH_3)_4]^{2-}$ and $[Fe(SCH_3)_3]^{1-}$ sites were slightly more sensitive. For B3LYP, the full core 6-31G** and DZVP2 basis sets gave comparable results for the Fe−S bond, but 6-31G** gave a slightly better decrease in the Fe−S bond length upon oxidation mainly because the DZVP2 Fe−S bond for $[Fe(SCH_3)_4]^{2-}$ is too long by ∼0.07 Å. Adding sp-type diffuse functions to the sulfur atoms changed the Fe−S bond lengths only slightly, while adding more polarization functions caused them to shorten, closer to experiment. Apparently, the radial flexibility obtained from the diffuse functions was less important than angular flexibility from the additional polarization functions for obtaining the correct geometries. Moreover, of the other internal angles, the S−C bond (Table S2) shortens by ∼0.015 Å between ECPDZ and ECPDZ+DZVP where the full core replaces the ECP on just the ligand atoms but changes insignificantly on going to either the full core DZVP2 or the 6-31G** basis sets; thus, ECPDZ+DZVP is a reasonable compromise for the optimization of larger molecular systems. Also, the B3LYP/DZVP2 S−C bond for $[Fe(SCH_3)_4]^{2-}$ is about 0.015 Å long, which is a greater error than for B3LYP/6-31G** S−C bonds of any of the analogues.

Overall, the B3LYP method using ECPDZ+DZVP, 6-31G**, and 6-31G(3df)$_S$** as small, medium, and slightly larger basis sets, respectively, gives good geometries with respect to the size of calculation. The geometries for these calculations are summarized in Table 1 with results for the others in Tables S1−S5. The B3LYP method using 6-31(++)$_S$G**, DZVP2, and DZVP2(++)$_S$ also gave good geometries. However, the addition of the diffuse functions for both 6-31G** and DZVP2 basis sets increased the size of the calculation without significantly changing the results. Also, although the 6-31G** and DZVP2 basis sets behaved similarly and are of similar size, DZVP2 basis sets gave slightly poorer geometries for $[Fe(SCH_3)_4]^{2-}$, and the 6-31G** basis sets are widely used and have been extended to include sp-type diffuse functions (6-31++G**) and additional polarization functions [6-31G(3df,3pd)].

Dependence of Oxidation Energies of $[Fe(SCH_3)_4]^{1-/0}$ and $[Fe(SCH_3)_3]^{1-/0}$ on Electron Correlation, Basis Set, and Density Functional. To assess functional and basis set effects on energetics, the ADE and VDE of $[Fe(SCH_3)_4]^{1-}$ and $[Fe(SCH_3)_3]^{1-}$ were calculated using energies without ZPE corrections obtained at the B3LYP and B97gga1 levels with the ECPDZ+DZVP, 6-31G**, 6-31(++)$_S$G**, and

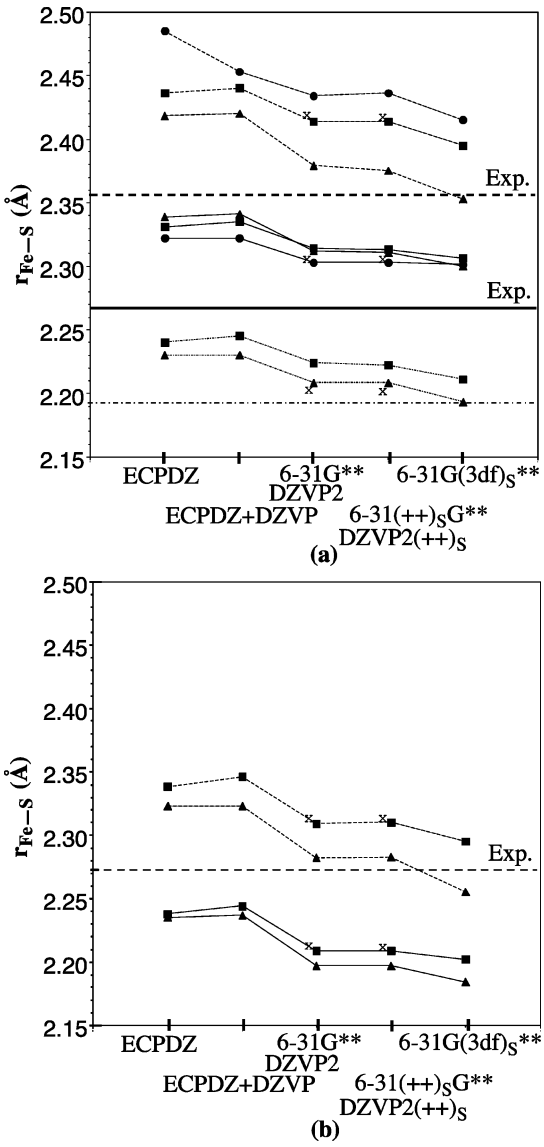


Figure 2. Optimized Fe−S bond lengths at the BHandH (●), B3LYP (■), and B97 (▲) levels as a function of basis sets. In addition, B3LYP/DZVP2 and DZVP2(++)$_S$ are shown (×). (a) $[Fe(SCH_3)_4]^{2-}$ (- - -), $[Fe(SCH_3)_4]^{1-}$ (−), and $[Fe(SCH_3)_4]^0$ (- • - • -) with experimental values indicated for $[Fe(S_2\text{-}o\text{-xyl})_2]^{2-}$ (- - -), $[Fe(S_2\text{-}o\text{-xyl})_2]^{1-}$ (−), and [Fe(1,2-benzenedithiolate)$_2$(PMe$_3$)]^0 (- • - • -). (b) $[Fe(SCH_3)_3]^{1-}$ (- - -) and $[Fe(SCH_3)_3]^0$ (−) with experimental values indicated for $[Fe(SC_6H_2\text{-}2,4,6\text{-}t\text{-}Bu_3)_3]^{1-}$ (- - -).

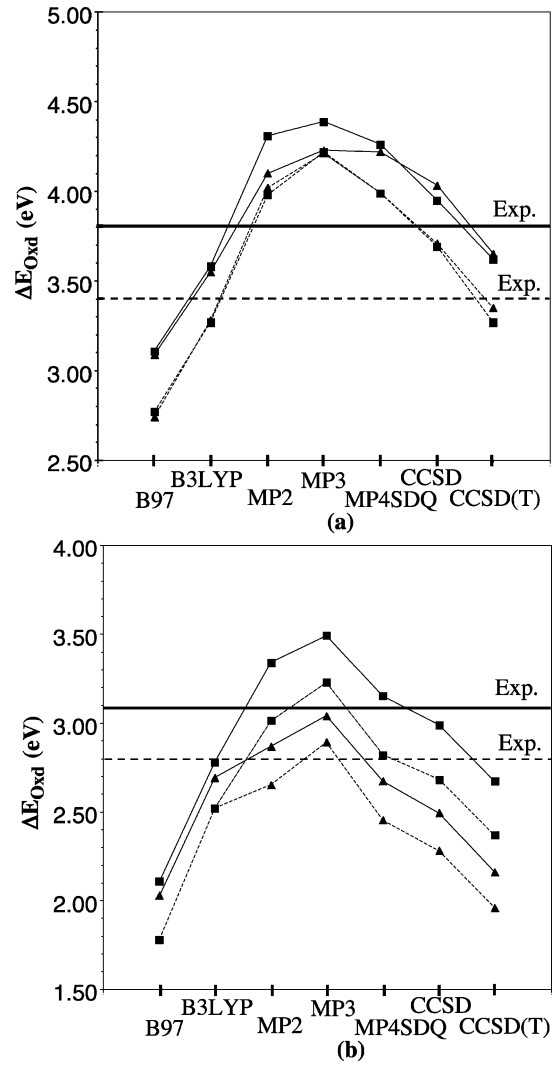


Figure 3. The calculated VDE (−) and ADE (- - -) using the 6-31G** basis as a function of the levels of calculation for the B3LYP/6-31G** (■) and B97gga1/6-31G** (▲) geometries. (a) $[Fe(SCH_3)_4]^{1-}$ with experimental values indicated for $[Fe(SCH_2CH_3)_4]^{1-}$ and (b) $[Fe(SCH_3)_3]^{1-}$ with experimental values indicated for $[Fe(SCH_3)_3]^{1-}$.

6-31G(3df)$_S$** basis sets (Tables S6 and S8), using geometries calculated at various levels. The energies were compared to PES results for ADE and VDE of $[Fe(SCH_3)_4]^{1-/0}$ and $[Fe(SCH_3)_3]^{1-/0}$ [16,17] and also to energies calculated at the MPn and coupled cluster levels.

First, the general performance of the B3LYP and B97gga1 functionals for calculating ADE and VDE against experiment and the MP2, MP3, MP4SDQ, CCSD, and CCSD(T) levels of theory was evaluated for the B3LYP/6-31G** and B97gga1/6-31G** geometries (Figure 3, Table S6). The energies were evaluated using the 6-31G** basis so that a consistent basis set could be used even for the very large coupled cluster calculations. The energies were also calculated for the B3LYP/ECPDZ+DZVP geometries (Table S6), but the results showed trends similar to those of the B3LYP/6-31G** geometries. Although the B97gga1 optimized geometries were in good agreement with experimental values, the ADE and VDE calculated from B97gga1 energies were underestimated by ∼0.7 to ∼1 eV as compared to experiment, while they were generally much better at the B3LYP level where the energies were underestimated by up to ∼0.3 eV as compared to experiment. Because B97gga1 energy calculations using either B3LYP/6-31G** or B97gga1/6-31G** geometries gave very similar results, this is not due to the poor changes in Fe−S bond length upon reduction in B97gga1 as compared to the B3LYP but rather a poorer treatment of the energy. Examining the other methods, the MP series failed as an appropriate method to include correlation energy especially for $[Fe(SCH_3)_3]^{1-}$, while the coupled cluster methods generally gave better predictions; however, the CCSD(T) ADE and VDE of $[Fe(SCH_3)_3]^{1-}$ were overestimated by about 0.4 eV apparently due to degeneracy or near-degeneracy problems. Overall, B3LYP/

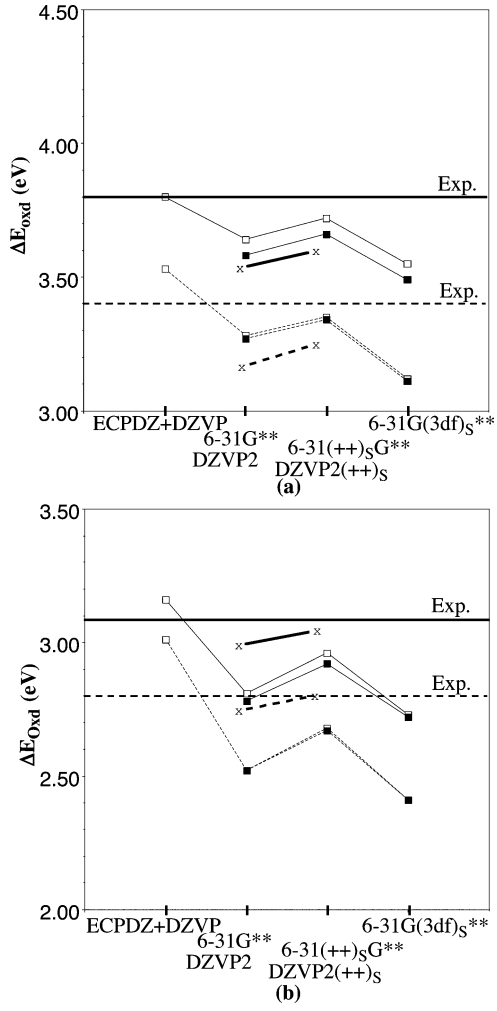


Figure 4. The calculated VDE (−) and ADE (- - -) as a function of basis sets at the B3LYP level for the B3LYP/ECPDZ+DZVP (□) and B3LYP/6-31G** (■) geometries. Results plotted at B3LYP/DZVP2 and B3LYP/DZVP2(++)$_S$ (×) used the B3LYP/DZVP2 and B3LYP/DZVP2(++)$_S$ geometries, respectively. (a) $[Fe(SCH_3)_4]^{1-}$ with experimental values indicated for $[Fe(SCH_2CH_3)_4]^{1-}$ and (b) $[Fe(SCH_3)_3]^{1-}$ with experimental values indicated for $[Fe(SCH_3)_3]^{1-}$.

6-31G** gave much better oxidation energies than B97gga1, which equaled the accuracy of the energy calculations at the very costly CCSD level.

Next, the performance of different basis sets for calculating ADE and VDE against experiment was evaluated using the B3LYP functional with both the B3LYP/ECPDZ+DZVP and the B3LYP/6-31G** geometries (Figure 4, Table S8). Results for B3LYP/DZVP2 and B3LYP/DZVP2(++)$_S$ calculations of geometry and energy also shown for comparison. The performance was also evaluated for the B97gga1 functional with the B97gga1/ECPDZ+DZVP and B97gga1/6-31G** geometries (Table S8); however, the overall energies were underestimated by up to ∼1 eV as indicated above, and the trends found for B3LYP with increasing basis set size were the same. The calculated ADE and VDE from B3LYP tend to decrease as the basis set size increases from the ECPDZ+DZVP to full core basis sets because larger basis sets can more accurately approximate the orbitals by imposing fewer restrictions on the spatial distribution of the electrons and by describing the defuse nature of anions.[44,45] However, relative to the 6-31G** basis sets, adding diffuse functions to the sulfurs increases and thus improves the calculated ADE and VDE with respect to experiment, while adding more polarization functions to the sulfurs decreases and thus worsens these values. In particular, while addition of diffuse functions or additional polarization functions to the sulfurs lowers the energy for both the oxidized and the reduced analogues by ∼0.1 to 0.2 eV and ∼1.2 to 1.5 eV, respectively, the diffuse functions lower the energy of the anionic reduced analogue by ∼0.1 eV more than that of the neutral oxidized analogue, while the additional polarization functions lower the energy of the neutral analogue by ∼0.1 eV more than the anionic analogue (Table S9). Other PES and DFT studies on the tetrahedral ferric complexes $Fe^{III}X_4^-$ (X = Cl, Br) and the three-coordinate complexes $M^{II}X_3^-$ (M = Mn, Fe, Co, Ni; X = Cl, Br) also indicate that increasing the basis set size for all atoms, for example using triple-ζ basis sets, does not significantly change the calculated geometric parameters and redox energies.[46]

Overall, the B3LYP method using ECPDZ+DZVP, 6-31G**, and 6-31(++)$_S$G**, as small, medium, and slightly larger basis sets, respectively, gives good ADE and VDE using the B3LYP ECPDZ+DZVP and 6-31G** geometries. The ADE and VDE for these calculations are summarized in Table 2 with results for the other calculations in Tables S6 and S8. The B3LYP method using the DZVP2 and DZVP2(++)$_S$ basis sets also gave good ADE and VDE, but as mentioned previously, the 6-31G** are widely used and have been extended to include sp-type diffuse functions and additional polarization functions. However, energies using 6-31G(3df)$_S$** are in even poorer agreement than energies using ECPDZ+DZVP at a much greater computational cost, so adding diffuse functions is preferred to adding more polarization functions. Also, the energy

Table 2. Experimental and Calculated ADE and VDE (in eV) for $[Fe(SCH_3)_4]^-$ and $[Fe(SCH_3)_3]^-$ at the B3LYP Level of Theory Using Different Basis Sets

	$[Fe(SCH_3)_4]^-$		$[Fe(SCH_3)_3]^-$	
method[a]	ADE	VDE	ADE	VDE
B3LYP/ECPDZ+DZVP	3.53	3.80	3.01	3.16
B3LYP/6-31G**//G(1)	3.28	3.64	2.52	2.81
B3LYP/6-31(++)$_S$G**//G(1)	3.35	3.72	2.68	2.96
B3LYP/6-31G**	3.27	3.58	2.52	2.78
B3LYP/6-31(++)$_S$G**//G(2)	3.34	3.66	2.67	2.92
B3LYP/6-31(++)$_S$G**//G(2)[c]	3.35	3.62	2.77	2.98
exp.[b]	3.40 (±0.06)	3.80 (±0.06)	2.80 (±0.06)	3.08 (±0.06)

[a] G(1), B3LYP/ECPDZ+DZVP geometry; G(2), B3LYP/6-31G** geometry. [b] References 16 and 17. [c] With the ZPE correction.

calculations at the B3LYP level are not very sensitive to the geometries tested here so that reasonable reduction potential energies can be obtained based on optimized geometries with smaller basis sets. In particular, the B3LYP/6-31(++)$_S$G** energies for the 6-31G** and 6-31(++)$_S$G** geometries (i.e., B3LYP/6-31(++)$_S$G**// B3LYP/6-31G** versus B3LYP/6-31(++)$_S$G** in Table S8) are within less than 0.01 eV so that optimization of the geometry can be done using the 6-31G** basis sets.

Conclusions

To find efficient computational methods for structural and energetic properties of iron−sulfur protein redox sites, spin-unrestricted B3LYP, BHandH, B97gga1, MP theory series, CCSD, and CCSD(T) calculations using a variety of ECP and full core basis sets have been performed on analogues of the rubredoxin redox site and compared to the experimental geometries, ADE, and VDE. Overall, the B3LYP method gives the best results in comparison with experiment and is comparable for energies to those at the very costly CCSD or CCSD(T) level of theory. For geometries, B3LYP using ECPDZ+DZVP basis sets overestimated the Fe−S bonds by ∼0.06 to ∼0.08 Å, using 6-31G** by ∼0.03 to ∼0.06 Å, and using 6-31G(3df)$_S$** by ∼0.02 to ∼0.04 Å, but all gave changes in Fe−S bond length upon oxidation within ∼0.01 Å of experiment and much more accurate S−C bonds. The energies are relatively insensitive to geometries, giving similar results for the B3LYP/ECPDZ+DZVP and B3LYP/6-31G** geometries. For energy calculations, the B3LYP functional using ECPDZ+DZVP basis sets overestimates by no more than ∼0.2 eV of experiment, using 6-31G** underestimates by no more than ∼0.3 eV, and using 6-31(++)$_S$G** underestimates by no more than ∼0.15 eV. Adding zero point energy corrections improves the 6-31(++)$_S$G** energies so that the ADE are underestimated by no more than 0.05 eV, although the VDE for $[Fe(SCH_3)_3]^{1-}$ is underestimated by 0.18 eV.

In summary, a central issue in the reduction potential calculations of iron−sulfur systems is to obtain relatively accurate energies for both the oxidized and the reduced states. The recommended calculation for iron−sulfur systems is to optimize the geometry using B3LYP with at least a double-ζ basis set with polarization functions such as 6-31G** and DZVP2 and then to calculate energies using B3LYP with additional diffuse functions such as from 6-31++G** added to the sulfurs [6-31(++)$_S$G** and DZVP2(++)$_S$]. However, for larger iron−sulfur systems in QM/MM calculations, a reasonable calculation would be to optimize the geometry using B3LYP with an ECP for iron and then to calculate energies calculated using B3LYP/6-31(++)$_S$G** or DZVP2-(++)$_S$.

Acknowledgment. This work was supported by a grant from the National Institutes of Health (GM45303) and by a grant (GC3565 and GC20901) from the Molecular Science Computing Facility (MSCF) in the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL), a national user facility sponsored by the U.S. DOE's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory, operated for DOE by Battelle. We thank Dr. David Dixon, Prof. Lai-Sheng Wang, and Dr. Xubin Wang for helpful discussions. T.I. thanks Dr. Bernard R. Brooks at the NIH for his hospitality during part of these studies.

Supporting Information Available: Supplemental tables (S1−S9) of the optimized geometries and calculated energies. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Holm, R. H. In *Advances in Inorganic Chemistry*; Cammack, R., Ed.; Academic Press: San Diego, 1992; Vol. 38, pp 1−71.

(2) Beinert, H.; Holm, R. H.; Münck, E. *Science* **1997**, *277*, 263–659.

(3) Rao, P. V.; Holm, R. H. *Chem. Rev.* **2004**, *104*, 527–559.

(4) Beinert, H. *J. Biol. Inorg. Chem.* **2000**, *5*, 2–15.

(5) Spiro, T. G. *Iron-Sulfur Proteins*; Wiley: New York, 1982.

(6) Ichiye, T. In *Computational Biochemistry and Biophysics*; Becker, O. M., MacKerell, A. D., Jr., Roux, B., Watanabe, M., Eds.; Marcel Dekker, Inc.: New York, 2001; pp 393−415.

(7) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: Oxford, 1989.

(8) Niu, S. Q.; Hall, M. B. *Chem. Rev.* **2000**, *100*, 353–405.

(9) Noodleman, L.; Case, D. A. In *Advances in Inorganic Chemistry*; Cammack, R., Ed.; Academic Press: San Diego, 1992; Vol. 38, pp 423−470.

(10) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J. M. *Coord. Chem. Rev.* **1995**, *144*, 199–244.

(11) Noodleman, L.; Lovell, T.; Han, W. G.; Li, J.; Himo, F. *Chem. Rev.* **2004**, *104*, 459–508.

(12) Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2000**, *316*, 160–166.

(13) Li, J. *Acta Chim. Sin.* **2000**, *58*, 1529–1533.

(14) Sigfridsson, E.; Olsson, M. H. M.; Ryde, U. *Inorg. Chem.* **2001**, *40*, 2509–2519.

(15) Szilagyi, R. K.; Winslow, M. A. *J. Comput. Chem.* **2006**, *27*, 1385–1397.

(16) Wang, X. B.; Wang, L. S. *J. Chem. Phys.* **2000**, *112*, 6959–6962.

(17) Yang, X.; Wang, X. B.; Fu, Y. J.; Wang, L. S. *J. Phys. Chem. A* **2003**, *107*, 1703–1709.

(18) Niu, S.-Q.; Wang, X.-B.; Nichols, J. A.; Wang, L.-S.; Ichiye, T. *J. Phys. Chem. A* **2003**, *107*, 2898–2907.

(19) Wang, X.-B.; Niu, S.-Q.; Yang, X.; Ibrahim, S. K.; Pickett, C. J.; Ichiye, T.; Wang, L.-S. *J. Am. Chem. Soc.* **2003**, *125*, 14072–14081.

(20) Yang, X.; Niu, S.; Ichiye, T.; Wang, L.-S. *J. Am. Chem. Soc.* **2004**, *126*, 15790–15794.

(21) Niu, S.-Q.; Ichiye, T. *Theor. Chim. Acta* **2007**, *117*, 275–281.

(22) Perrin, B. S., Jr.; Ichiye, T. Unpublished work.

(23) Moura, I.; Pereira, A. S.; Tavares, P.; Moura, J. J. G. In *Iron-Sulfur Proteins*; Sykes, A. G., Cammack, R., Eds.; Academic Press, Inc.: San Diego, 1999; Vol. 47, pp 361−419.

(24) Macdonnell, F. M.; Ruhlandtsenge, K.; Ellison, J. J.; Holm, R. H.; Power, P. P. *Inorg. Chem.* **1995**, *34*, 1815–1822.

(25) Sellmann, D.; Geck, M.; Knoch, F.; Ritter, G.; Dengler, J. *J. Am. Chem. Soc.* **1991**, *113*, 3819–3828.

(26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(27) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785–789.

(28) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554–8560.

(29) Couty, M.; Hall, M. B. *J. Comput. Chem.* **1996**, *17*, 1359–1370.

(30) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.

(31) Stevens, W. J.; Basch, H.; Krauss, M. *J. Chem. Phys.* **1984**, *81*, 6026–6033.

(32) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293–1302.

(33) Dunning, T. H., Jr. *J. Chem. Phys.* **1970**, *53*, 2823–2833.

(34) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560–571.

(35) Francl, M. M.; Petro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(36) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222.

(37) Clark, T.; Spitznagel, G. W.; Chandrasekhar, J.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294–301.

(38) Levine, I. N. *Quantum Chemistry*, 5th ed.; Prentice Hall: Upper Saddle River, NJ, 2000.

(39) 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.; V., T.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtl, 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.; 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.; 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 4.6*; Pacific Northwest National Laboratory: Richland, WA, 2004.

(40) 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 C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(41) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H. *Orbital Interactions in Chemistry*; John Wiley and Sons: New York, 1985.

(42) Lane, R. W.; Ibers, J. A.; Frankel, R. B.; Papaefthymiou, G. C.; Holm, R. H. *J. Am. Chem. Soc.* **1977**, *99*, 84–98.

(43) Maelia, L. E.; Millar, M.; Koch, S. A. *Inorg. Chem.* **1992**, *31*, 4594–4600.

(44) Foresman, J. B.; Frisch, A. E. *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1996.

(45) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; John Wiley & Sons: New York, 1986.

(46) Yang, X.; Wang, X.-B.; Wang, L.-S.; Niu, S.; Ichiye, T. *J. Chem. Phys.* **2003**, *119*, 8311–8320.

CT800357C

JCTC Journal of Chemical Theory and Computation

Electro-Optical Parameters for Computation of Nonresonance Raman Scattering Intensities of Peptides

Vineet Gupta,[†] Konstantin S. Smirnov,*[,‡] Daniel Bougeard,[‡] and Poonam Tandon[†]

Department of Physics, Lucknow University, 226007 Lucknow, India, and LASIR, Université des Sciences et Technologies de Lille, CNRS, Bât. C5, 59655 Villeneuve d'Ascq, France

Received November 26, 2008

Abstract: A set of electro-optical parameters (EOPs) of cylindrical zero-order bond polarizability model (BPM) for chemical bonds found in peptides was obtained from the results of quantum-chemical computations. The calculation of the polarizability tensors and the Raman scattering activities of four test molecules (N-methylacetamide, N,N-dimethylacetamide, dialanine, and diglycine) has shown that the BPM calculated quantities are in good agreement with the reference data obtained in experiments or derived by quantum-chemical calculations. The mean molecular polarizabilities are reproduced with the maximum relative error of 1.6%. A good agreement was obtained for the Raman activity of stretching vibrations, whereas a limited performance of the EOPs was found for the vibrational modes with a significant contribution of the bending internal coordinates involving H atoms. The origins of the discrepancies are analyzed, and the ways of improvement of the model performance are discussed.

1. Introduction

Response of electronic structure of molecules to an external electric field and the dependence of the response on a variation of molecular geometry play a fundamental role in reactivity and optical activity of molecules. For an electric field of low intensity the response is determined by the dipole polarizability tensor, whereas the change of the tensor with the geometry of the molecule is defined by the derivatives of the tensor with respect to displacements of nuclei. The latter quantities are directly related to the intensities in the Raman scattering spectra of the molecule. Raman spectroscopy has recommended itself as a fast, reliable, and nondestructive means of studying the structure sensitive dynamics of systems in the time scale from femtoseconds to picoseconds.

The calculation of the polarizability and the quantitative modeling of the Raman scattering intensities of peptides and proteins can be of significant help for understanding the relationship between structure, function, and dynamics of these (macro-)molecules.[1,2] Furthermore, such modeling can be of significant help for developing both structural and potential models of these systems. The simulation of spectral signatures of the vibrational dynamics is often limited to the computation of the density of vibrational states, that is a poor indicator of the validity of the models since for systems with many atoms the densities of vibrational states computed with different potential models strongly resemble each other because of the absence of any selection rules.[3] On the other hand, the infrared and Raman spectra in which the modes are discriminated according to selection rules often provide more valuable information, but examples of the IR and Raman spectra calculations are scarce, particularly because of the lack of intensity parameters.

The polarizability tensor as well as its derivatives with respect to atomic displacements can be computed in different ways. The modern quantum-chemical methods permit the evaluation of these quantities with a high precision, but such calculations are still fairly limited to systems of small size, especially if the derivatives of the polarizability tensor have to be computed. For real systems containing hundreds/thousands of atoms the only feasible way is to use parametric models that permit a fast calculation of the quantities of interest. Several such models exist[4−8] among which the bond

* Corresponding author phone: +33 320336139; fax: +33 320436755; e-mail: Konstantin.Smirnov@univlille1.fr.

[†] Lucknow University.

[‡] Université des Sciences et Technologies de Lille.

10.1021/ct800510y CCC: $40.75
Published on Web 04/08/2009

polarizability model[4] (BPM) permits a fast and reliable approach for computing both the polarizability and the Raman scattering intensities.

The BPM uses a set of parameters (electro-optical parameters, EOPs) to construct, via the topology of chemical bonds and the geometry of the system, the polarizability tensor and to describe the change of the polarizability with geometry variations. As in any parametric model, the quality of the parameters determines the reliability and predictive power of the model. Experimentally, EOPs can be extracted from data of Raman scattering experiments, but the resulting values of the parameters are often ambiguous because the number of observables is generally smaller than the number of EOPs to be determined. The ill-conditioned mathematical problem as well as the fact that measurements of the absolute Raman intensities are necessary to obtain absolute values of parameters has as a consequence the fact that sets of EOPs were determined only for few systems, e.g. alkanes,[9] while most of the works employing the BPM use only relative values of the parameters. It is therefore highly desirable to have a way of deriving reliable sets of EOPs whose values can be interpreted with a physical sense but not as solution of a mathematical problem.

Recently, an approach for obtaining EOPs of BPM from results of *ab initio* quantum-chemical calculations of molecular models was proposed.[10] The advantage of the procedure is that it overcomes the two problems mentioned above because the calculations are in principle capable of providing the necessary observables. The sets of EOPs derived in this way for systems with different types of chemical bonds (alkanes[10] and aluminosilicates[11]) were shown to be in good agreement with available experimental and theoretical data, although the EOPs for aluminosilicates with partially ionic bonds were found to have a limited transferability to periodic systems.[11]

The present paper reports the development of a set of the BPM electro-optical parameters for chemical bonds relevant to molecules of biological interest, in particular to peptides. The paper is organized as follows. The next section (section 2) discusses the BPM and the procedure used to derive the EOPs. This discussion is followed by a presentation of the molecular models of the training and validation sets and by details of quantum-chemical calculations whose results are used as reference data in the development. Results of the development and the analysis of discrepancies and of limitations of the model are given in section 3, which also presents the application of the obtained set of EOPs to the calculation of the polarizability and of the Raman spectra of molecules in the validation set. The conclusions of the work are summarized in section 4.

2. Model and Computational Procedures

2.1. Bond Polarizability Model. This section provides a brief description of BPM; the interested reader can find a thorough presentation of BPM in refs 9, 12, and 13. The BPM represents the polarizability tensor **A** of a molecule as the sum of polarizability tensors of the bonds in the molecule

$$\mathbf{A} = \sum_{i \in \text{bonds}} \mathbf{a}_i \tag{1}$$

with $\mathbf{a}_i$ being the polarizability tensor of bond i in the fixed Cartesian frame. The tensors a_i can be rewritten as

$$\boldsymbol{\alpha}_i = \mathbf{R}_i \alpha_i \mathbf{R}_i^{-1} \tag{2}$$

where $\boldsymbol{\alpha}_i$ is the bond polarizability tensor in its principal axes, and $\mathbf{R}_i$ is the rotation matrix between the bond principal and the Cartesian frames. The three nonzero components of the α_i tensor are the longitudinal (L) and two transversal (T and T') bond polarizabilities. The polarizability tensor at the equilibrium $\mathbf{A}^0$ is therefore equal to

$$\mathbf{A}^0 = \sum_{i \in \text{bonds}} \mathbf{R}_i \boldsymbol{\alpha}_i^0 \mathbf{R}_i^{-1} \tag{3}$$

and the Cartesian components of the $\mathbf{A}^0$ tensor can be written as

$$A_{pq}^0 = \sum_d \sum_b \alpha_d^0(b) \sum_{i \in b} p_d(i) q_d(i) \tag{4}$$

where $p, q = x, y, z$, while the indexes d and b run over the principal components ($d = L, T, T'$) of the bond polarizability tensor $\boldsymbol{\alpha}_i$ and bond types, respectively.

The intensity I_k of the peak corresponding to the vibrational mode k in the Raman spectrum of molecule reads[14]

$$I_k \propto \frac{\omega_s^4}{\omega_k(1 - \exp(-\beta\hbar\omega_k))} \left\{ \frac{45(\bar{\alpha}')^2 + 7(\gamma')^2}{45} \right\} \tag{5}$$

where ω_k and ω_s are the angular frequencies of the mode k and of the scattered radiation, respectively, and β and $\hbar$ have their usual meaning. The quantities $\bar{\alpha}'$ and γ' in eq 5 stand for the trace and anisotropy of the polarizability tensor derivative $\partial\mathbf{A}/\partial Q_k$ with respect to the normal coordinate Q of mode k; these are calculated in the usual way[14]

$$\bar{\alpha}' = \frac{1}{3}(A'_{xx} + A'_{yy} + A'_{zz}) \tag{6}$$

$$(\gamma')^2 = \frac{1}{2}[(A'_{xx} - A'_{yy})^2 + (A'_{yy} - A'_{zz})^2 + (A'_{zz} - A'_{xx})^2 + 6(A'^2_{xy} + A'^2_{xz} + A'^2_{yz})] \tag{7}$$

with $A'_{pq} \equiv \partial A_{pq}/\partial Q_k$. Note that eq 5 was obtained assuming a linearly polarized incident radiation.

The Raman scattering intensity I_k of the mode k is therefore determined by the tensor $\mathbf{A}'_k$ that can be written as

$$\mathbf{A}'_k \equiv \frac{\partial \mathbf{A}}{\partial Q_k} = \sum_{i \in \text{bonds}} \left\{ \frac{\partial \mathbf{R}_i}{\partial Q_k} \boldsymbol{\alpha}_i^0 \mathbf{R}_i^{-1} + \mathbf{R}_i \boldsymbol{\alpha}_i^0 \frac{\partial \mathbf{R}_i^{-1}}{\partial Q_k} + \mathbf{R}_i \left[\sum_j \boldsymbol{\alpha}'_{ij} \left(\frac{\partial s_j}{\partial Q_k} \right) \right] \mathbf{R}_i^{-1} \right\} \tag{8}$$

where the index j runs over the internal coordinates s_j. The first two terms in eq 8 account for the variation of the molecular polarizability with changes of bond orientations in the normal mode Q_k, while the last term explains the change of the bond polarizability with the variation of internal coordinates. The $\boldsymbol{\alpha}_i^0$ and $\boldsymbol{\alpha}'_{ij} \equiv (\boldsymbol{\alpha}_i/\partial s_j)_0$ quantities in eq 8 are known as equilibrium and valence electro-optical parameters of the bond polarizability model, respectively.

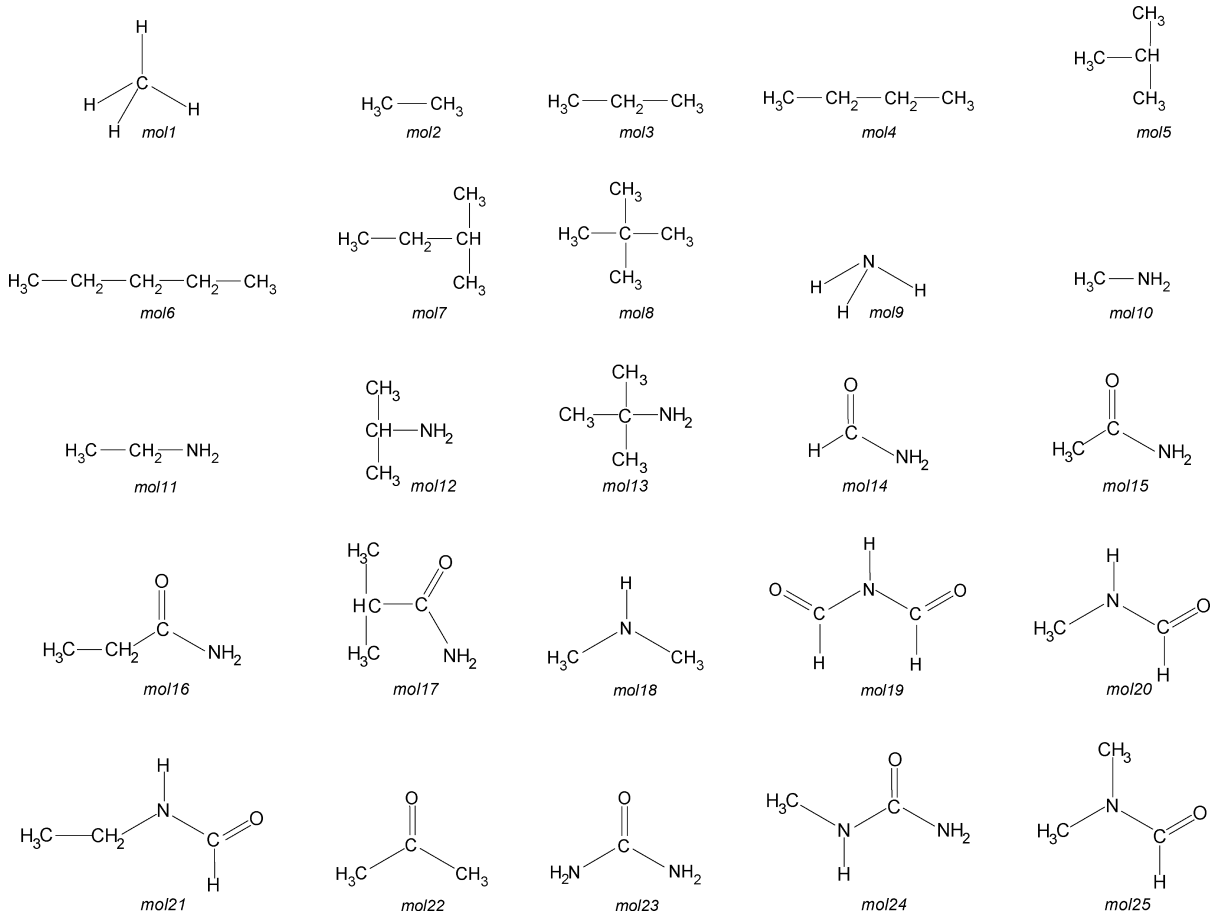


Figure 1. Molecules of the training set used to derive electro-optical parameters.

In a similar way as eq 4 the components of the $\mathbf{A}'_k$ tensor can be represented as a linear combination of the EOPs with coefficients depending on the molecular geometry (matrices $\mathbf{R}_i$) and on the variations of the internal coordinates in the normal modes (derivatives $\partial s_j/\partial Q_k$). All these quantities can be obtained from the molecular geometry and from the Cartesian displacements of atoms in the normal mode k. Hence, for a molecule containing N atoms the eqs 3 and 8 form a system of $6 \times (3N - 5)$ linear equations with EOPs as unknowns.

In the following we shall limit ourselves to a cylindrical zero-order BPM which implies that (i) all but $\boldsymbol{\alpha}'_{ii}$ parameters are equal to zero (zero-order model) and that (ii) in the tensor $\boldsymbol{\alpha}_i^0$ and in the tensor $\boldsymbol{\alpha}'_{ii}$ the two transversal components are equal to each other (cylindrical model). The former assumption reduces eq 8 to

$$\mathbf{A}'_k = \sum_{i \in \text{bonds}} \left\{ \frac{\partial \mathbf{R}_i}{\partial Q_k} \boldsymbol{\alpha}_i^0 \mathbf{R}_i^{-1} + \mathbf{R}_i \boldsymbol{\alpha}_i^0 \frac{\partial \mathbf{R}_i^{-1}}{\partial Q_k} + \mathbf{R}_i \left[\boldsymbol{\alpha}'_{ii} \left(\frac{\partial s_i}{\partial Q_k} \right) \right] \mathbf{R}_i^{-1} \right\} \tag{9}$$

while the latter limits the number of EOPs to four per bond type. A model that retains terms with $j \neq i$ in the last sum of eq 8 is known as the first-order BPM.

To derive EOPs with the approach described above one needs a database of the Cartesian components of the $\mathbf{A}^0$ and $\mathbf{A}'_k$ tensors for a set of molecules with known geometry. As such information is not generally available from experimental data, the database was generated with the use of quantum-chemical calculations as described in the next section. The $6 \times (3N - 5)$ linear equations for each molecule are then combined to form an extended set of equations that can be solved by a linear least-squares (LLSQ) method yielding statistically meaningful values of sought parameters. The proposed approach differs from that used previously in refs 10 and 11 in the way of creating the database. Instead of calculating the polarizability tensor derivatives with respect to bond lengths of individual bonds by the finite-differences procedure,[10,11] the present approach uses the derivatives of the tensor in normal modes obtained in an analytical way. As the derivatives determine the Raman scattering activity of the vibrational modes, it is therefore expected that the resulting EOPs will provide reliable Raman scattering intensities, while keeping a high precision in the calculation of the equilibrium polarizability.

2.2. Training Set. The training set of molecules used to obtain the A^0_{pq} and $\partial A_{pq}/\partial Q_k$ quantities consisted of eight linear and branched alkanes and of 17 small organic molecules with C−H, C−C, N−H, C−N, and C═O bonds (Figure 1). Previously, it was found that values of EOPs can be influenced by both the chemical environment of the bond and the position and orientation of the bond in the molecule.[15,16] Nevertheless, to keep the number of parameters in reasonable limits, it was supposed that all the bonds of the same type can be described by the same set of EOPs. The only exception was done for the C−N bonds which were

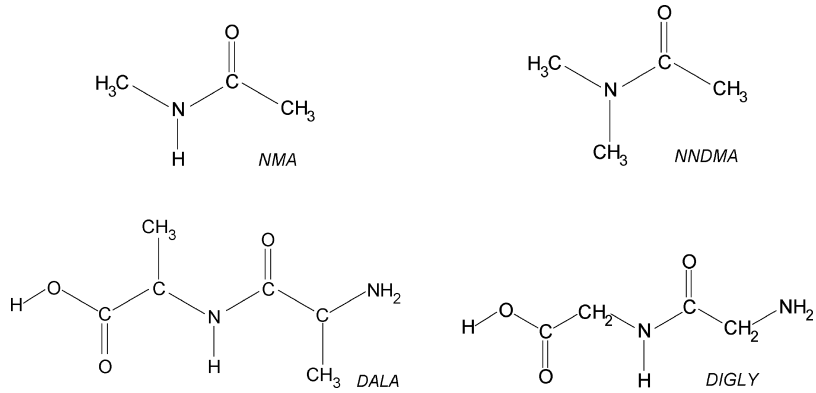


Figure 2. Molecules used in the validation set: N-methylacetamide (NMA), N,N-dimethylacetamide (NNDMA), dialanine (DALA), and diglycine (DIGLY).

separated into two types corresponding to C−N single and C−N partially double (peptide) bonds with bond lengths of *ca.* 1.45 Å and 1.36 Å, respectively. Therefore, six different types of bonds with a total number of 24 EOPs parameters were considered in the model.

For each molecule of the training set, the molecular geometry was optimized without symmetry constraints, and then the analysis of the vibrational normal modes was performed. The Cartesian displacements of atoms in the modes and the matrix of the Cartesian derivatives of the polarizability tensor produced in the calculations were then used to compute the $\partial A_{pq}/\partial Q_k$ quantities. The quantum-chemical calculations were carried out at the B3LYP level with the use of the pVTZ Sadlej basis set that was specially designed for the calculation of molecular electric properties, especially polarizability.[17,18] Halls and Schlegel have also shown that this basis set provides very good Raman activities for vibrational modes of small molecules.[19] The quantum-chemical calculations were performed with the G03 program.[20] The linear least-squares fitting of EOPs was done using the singular value decomposition method.[21]

The optimized geometry of the molecules of the training set is available as Supporting Information.

2.3. Validation Set. The derived set of EOPs was tested in the calculation of the polarizability tensors and Raman activities of N-methylacetamide (NMA), N,N-dimethylacetamide (NNDMA), dialanine (DALA), and diglycine (DIGLY) molecules (Figure 2) which can be considered as elementary building blocks of many peptide structures. Taking into account the fact that the experimental vibrational spectra of molecules in a condensed phase can substantially be affected by intermolecular interactions as well as reveal anharmonic effects (overtones and combination bands, Fermi resonance), the reference Raman activities R of the vibrational modes of the molecules were obtained in G03 calculations at the B3LYP/Sadlej basis set level after complete optimization of molecular geometries. These quantities were compared with those computed using the cylindrical zero-order BPM with obtained EOPs. In agreement with the definition used in the G03 program for the Raman activity the R_k quantities were calculated as

$$R_k = 45(\bar{\alpha}'_k)^2 + 7(\gamma'_k)^2 \qquad (10)$$

with $\bar{\alpha}'_k$ and γ'_k defined in eqs 6 and 7. One easily sees the relation of the $R_k$quantity with the intensity I_k in the Raman spectrum (eq 5). As the set of EOPs developed in the present work does not include C−O and O−H bonds, the calculations for the DALA and DIGLY molecules were done with the EOPs for C−O bond equal to those of C−N one, while the EOPs for the O−H bond were taken from ref 11.

3. Results and Discussion

Figure 3a,b presents the correlation graphs for the Cartesian components of the polarizability tensor A^0_{pq} and of the derivatives of the polarizability tensor A'_{pq} in the normal modes obtained in the quantum-chemical calculations and with the BPM for the molecules in the training set. The polarizability tensor components are, however, not usually measured quantities, and the response of a molecule to an applied electric field is commonly characterized by the mean molecular polarizability $\bar{\alpha} = (A_{xx} + A_{yy} + A_{zz})/3$. The BPM reproduces the mean polarizabilities of the 25 molecules of the training set with an rms error of 0.054 Å^3 and with the maximum relative error of 2.1% for the mean polarizability of NH_3 molecule. The good correlation seen in Figure 3a as well as the small discrepancy between the BPM computed and reference values of $\bar{\alpha}$ indicate that the equilibrium polarizability can be well described with the set of EOPs derived using the proposed approach. The good overall agreement seen in Figure 3b between the DFT and BPM computed A'_{pq} derivatives also implies that the set of EOPs obtained in the fitting procedure should yield correct Raman activities.

Figure 4a shows a correlation graph for the R_k values for the normal modes of the molecules in the training set. Although the accord between the DFT and BPM computed quantities can be considered as rather satisfactory, the inset in the figure reveals significant errors for some vibrational modes with a low activity. To identify the source of errors Figure 4b,c displays the correlation graphs for the $\bar{\alpha}'_k$ and γ'_k quantities, respectively, which determine the activity. One sees that a much better correlation exists for the trace of the polarizability tensor derivative than for the tensor's anisotropy.

The different extent of correlation in Figure 4b,c can in part be explained by different rms errors for the $\bar{\alpha}'_k$ and γ'_k

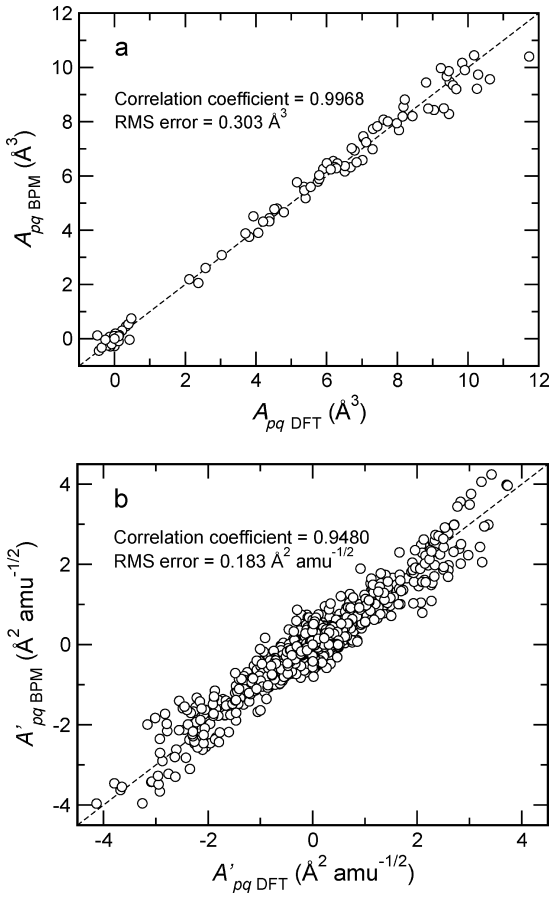


Figure 3. Correlation between the Cartesian components A_{pq} of the polarizability tensor (a) and between the components of the polarizability tensor derivatives $\partial A_{pq}/\partial Q_k$ (b) of molecules in the training set calculated by DFT and with BPM. Diagonal dashed line is the identity line.

derivatives. Assuming the Gaussian propagation rule, the variances σ^2 of these quantities can be estimated as

$$\sigma^2_{\bar{\alpha}'} \approx \frac{1}{3}\sigma^2_{A'} \tag{11}$$

$$\sigma^2_{\gamma'} \approx \left[1 + \frac{6(A'^{\,2}_{xy} + A'^{\,2}_{xz} + A'^{\,2}_{yz})}{(\gamma')^2}\right]\sigma^2_{A'} \tag{12}$$

where $\sigma^2_{A'}$ is the variance of the components of the $\mathbf{A}'$ tensor. One immediately sees that the rms error of the $\bar{\alpha}'$ derivative is smaller than that of A'_{pq} ones; the value of the error estimated to 0.106 Å^2 from eq 11 is in good agreement with that found for the fitted values and reported in Figure 4b. On the other hand, the rms error of the anisotropy derivative is always larger than the error of A'_{pq} quantities, and it lies in the range $\sigma_{A'}$ to $\sqrt{3}\sigma_{A'}$; for some vibrational modes the error can be comparable to the γ' value itself. This feature can result in a large error for the activity of modes with a high depolarization ratio whose intensity is largely determined by the γ'_k quantity. The above consideration however does not explain the scatter of points in Figure 4c that exceeds the estimated error for the γ'_k quantity, especialy for the values close to the graph's origin. This fact points to a deficiency of the used BPM model for the computation of the polarizability anisotropy derivative. It is worthy to note

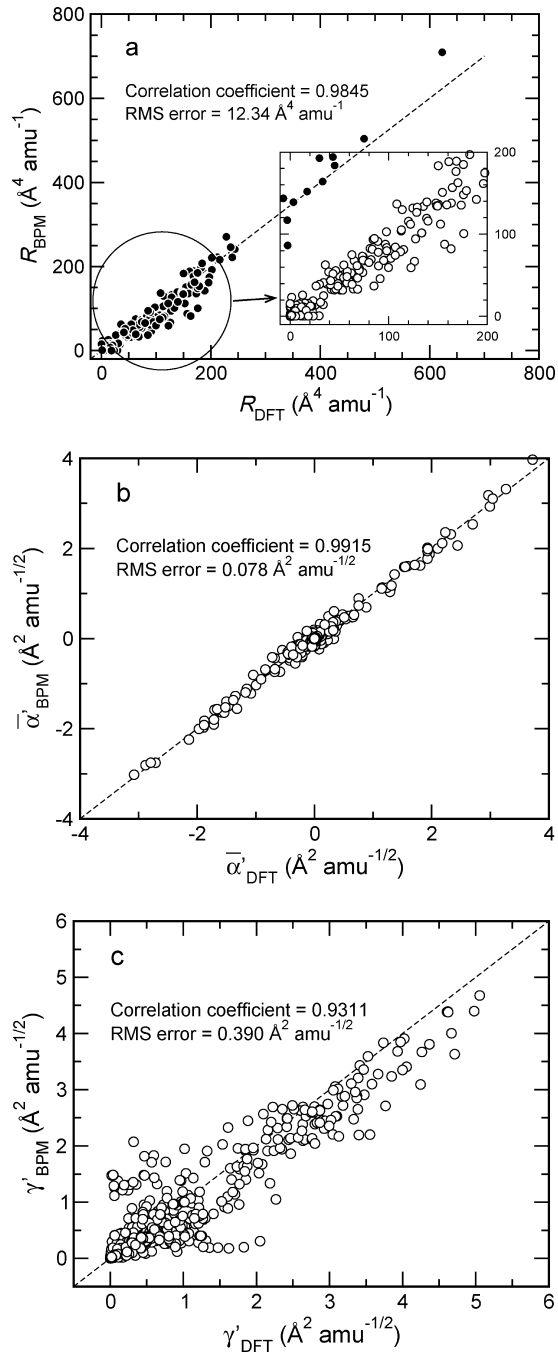


Figure 4. Correlation between the Raman activity (a), the derivative of mean polarizability $\bar{\alpha}'$ (b), and the anisotropy of the derivative of polarizability tensor γ' (c) calculated by DFT and with BPM for normal modes of molecules of the training set. Diagonal dashed line is the identity line.

that such a problem is inherent not only in the BPM but also in other models; this issue was already addressed in refs 22 and 23. An extension of the model to get around the deficiency is discussed below.

Table 1 reports the electro-optical parameters obtained in the LLSQ fit and compares them with those derived from the experimental Raman scattering intensities[24] or obtained in previous theoretical calculations.[10,25,26] A good agreement for parameters of different sets can only be found for the EOPs of C−H bond, while for other bonds the values of parameters vary significantly, particularly the longitudinal

Table 1. Values of Electro-Optical Parameters[f]

bond	equilibrium parameters (in Å^3) α_L^0	α_T^0	valence parameters (in Å^2) α'_L	α'_T	ref
C−H	*0.818 (0.092)*	*0.569 (0.087)*	*2.665 (0.090)*	*0.409 (0.061)*	
	0.779 (0.101)	0.489 (0.054)	2.743 (0.049)	0.353 (0.033)	10[e]
	0.82	0.50	2.14	0.44	24
	0.83	0.54	3.90	0.80	25
	0.83−0.88	0.53	2.43	0.45	26
N−H	*0.826 (0.141)*	*0.661 (0.126)*	*2.502 (0.196)*	*0.456 (0.133)*	
	1.50	0.50	1.0	0.24	24
C−C	*1.047 (0.302)*	*0.255 (0.273)*	*2.742 (0.493)*	*0.111 (0.349)*	
	1.677 (0.195)	0.127 (0.113)	2.863 (0.053)	0.243 (0.036)	10[e]
	0.7	0.50	3.0	0.5	24
	1.290−1.495	0.240−0.265	1.900	0.500−1.000	26
C−N_s[a]	*1.274 (0.389)*	*0.322 (0.328)*	*1.371 (0.810)*	*0.037 (0.561)*	
	0.49	0.44[d]	2.5	0.5	24
C−N_p[b]	*1.611 (0.444)*	*0.000*[c] *(0.393)*	*1.906 (0.866)*	*0.534 (0.597)*	
	0.28	0.49	2.0	0.50	24
C═O	*2.341 (0.402)*	*1.202 (0.352)*	*4.445 (0.824)*	*0.469 (0.576)*	
	2.56	1.04[d]	8.3	0.75[d]	24
	1.61	0.80	3.76	0.94	25, 26

[a] CN simple bond. [b] CN peptide bond. [c] The LLSQ fit resulted in value of α_T^0 parameter equal to −0.011 Å^3; the parameter was set to zero in the table. [d] Mean value of two transversal components. [e] See also Table 1 in ref 10 for a compilation of experimental values. [f] Statistical uncertainties are given in parentheses. Values given in italics refer to the results of the present work.

ones. There can be several reasons for such a discrepancy. First, it may reflect the ambiguity of EOPs derived on the basis of experimental Raman intensities. Analysis of Table 1 reveals that some parameters found in the literature have rather unexpected values thus indicating that the set may be one of several possible mathematical solutions of the problem. For example, one can expect values of N−H bond parameters in the range of those for C−H and O−H bonds, that is not the case for the EOPs obtained in ref 24. In contrast, the EOPs derived in the present work follow the expected trends. The present approach can therefore be of substantial help in choosing the solution that is meaningful from the physical viewpoint. Second, some differences in the parameters can be due to the use of a first-order BPM in refs 24−26 The use of the first-order BPM is, however, expected to influence the values of the principal parameters to a small extent.

The analysis of the standard errors of the EOPs in Table 1 shows that some parameters have a large uncertainty. The reason for this is 2-fold. First, not all chemical bonds are equally represented in the database; this can cause a larger uncertainty of EOPs for less frequently sampled bonds. Thus, the 25 molecules of the training set contain 148 C−H bonds as compared to 13 C−N_p bonds; note that the C−H bond EOPs have the smallest errors. Second, large uncertainty can point to a variation of the parameter while going from one molecule to another and in normal modes. Thus, Table 1 shows that valence EOPs of C−H and C−C bonds obtained in ref 10 generally have smaller uncertainties. These parameters were derived from the variations of molecular polarizability upon change of the lengths of *individual bonds* in molecules of training set, whereas the values of present EOPs were derived from the variation of the **A** tensor in *normal modes*. Large uncertainty of parameter can therefore be indicative of an important influence of other internal coordinates on the component of the bond polarizability tensor. The set of parameters reported in Table 1 can thus be considered as a generic set that may need further refinement for a particular class of molecules (at the expense of generality).

The set of EOPs was applied to the calculation of the polarizability and of the Raman scattering activity of the molecules of the validation set (Figure 2). Values of the components A_{pq}^0 of the polarizability tensor of the molecules are reproduced with an rms error of 0.869 Å^3, and a mean absolute error in the $\bar{\alpha}$ quantity of the four molecules is equal to 0.116 Å^3 with the maximum relative error of 1.6%. Note that the maximum discrepancies in the mean polarizability were obtained for the dialanine and diglycine molecules for which the values of EOPs for the C−O and O−H bonds of the carboxyl group were taken in a somewhat arbitrary way. The Raman activities R_k are reproduced with an rms error of 14.081 Å^4 amu^{-1} that is close to the rms for this quantity for the vibrational modes of the molecules in the training set.

Figures 5−8 compare the calculated Raman spectra of the molecules in the validation set. The spectra were obtained by the convolution of spectrum of Raman activities R_k with a Lorentzian function of 4 cm^{-1} width. One sees that intensities in the high-frequency region of reference spectra are well reproduced with the BPM. Given that the intensities are entirely due to C−H and N−H stretching vibrations, the good agreement between the BPM calculated and reference DFT spectra points to the fact that the valence EOPs for these bonds have reliable values. The same analysis concerns the mode at 1710−1730 cm^{-1} due to the stretching of the C═O bond (99% of the coordinate in potential energy distribution). The Raman activity of this mode is reproduced by the BPM model with a precision better than 10% thus indicating correct values of the valence parameters.

The most important differences between the spectra in the midfrequency range occur between 1200 and 1500 cm^{-1}; the BPM underestimates the peaks intensities in this region as compared to the DFT results. To our mind this fact points to the limits of the zero-order BPM. The vibrational modes in this energy range involve variation of many angle-bending

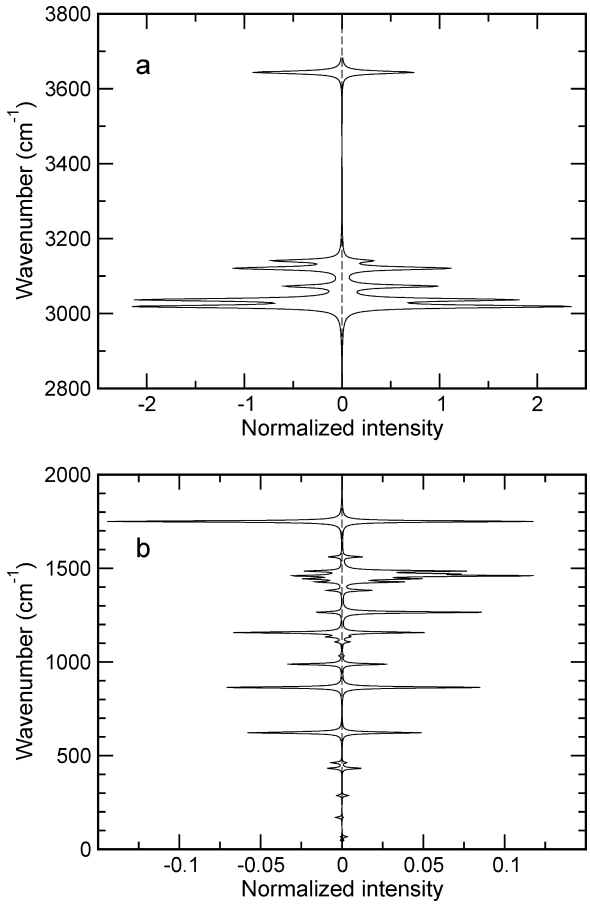


Figure 5. High-frequency (a) and low-frequency (b) parts of the Raman activity spectrum of N-methylacetamide molecule calculated with DFT (positive values of *x*-axis) and with BPM (negative values of *x*-axis). The intensity in the spectra were normalized on the total intensity and multiplied by 100; the intensities calculated with the BPM were made negative for comparison.

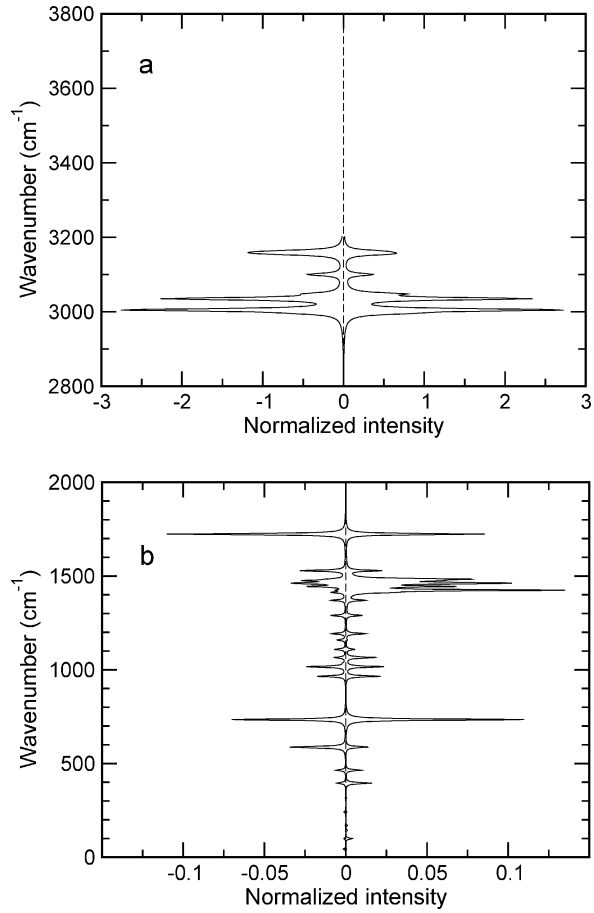


Figure 6. High-frequency (a) and low-frequency (b) parts of the Raman activity spectrum of N,N-dimethylacetamide molecule calculated with DFT (positive values of *x*-axis) and with BPM (negative values of *x*-axis). The intensity in the spectra were normalized on the total intensity and multiplied by 100; the intensities calculated with the BPM were made negative for comparison.

internal coordinates, particularly those including H-atoms. The zero-order BPM attempts to describe the related change in the molecular polarizability solely by contributions due to changes in bonds orientations, while it lacks the change induced in the bond polarizabilities upon variation of the bending coordinates. The missing polarizability variation could be simply described by a change of the equilibrium EOPs. However, such a modification will alter components of the equilibrium polarizability tensor, and, therefore, the improvement in the Raman scattering intensities will occur at a sacrifice of quality of the A^0_{pq} quantities.

The limited performance of BPM in describing the Raman intensity of bending modes has already been mentioned by Van Hemert and Blom.[23] In agreement with the exprimental work of Montero and Bermejo,[27] the authors of ref 23 pointed to the need of explicit consideration of parameters depending on the bending coordinates. As it was mentioned above, indirect support in the favor of such an extension is provided by large statistical uncertainties of some EOPs derived in the present work (Table 1). The extension would necessitate to keep in eq 8 α'_{ij} terms for some angle-bending internal coordinates s_j, that will, however, notably increase the number of EOPs.

Figure 9 compares the experimental Raman spectra of the N-methylacetamide and N,N-dimethylacetamide molecules in the gas phase[28] with those computed by eq 5 using the characteristics of experimental setup. To mimic the rotational broadening of Raman peaks in the experimental spectra, the theoretical ones were convoluted with a Lorentzian function with HWHM = 20 cm^{-1}. Besides the aforementioned discrepancy in the region of 1200−1500 cm^{-1}, the figure displays a very good agreement between the experimental and calculated spectra once more indicating that the EOPs can be used for the quantitative modeling of both the Raman scattering intensity and polarizability of peptide structures.

As the purpose of the EOPs model is the computation of large systems, we have also calculated the Raman spectrum of polyglycine I (PGI) by normal-mode analysis. The vibrational dynamics of this system has been the subject of many computational studies,[29−35] but the analysis of the dynamics in those works was limited to the calculation of the frequencies of vibrational modes and their dispersion curves, classification of modes according to the symmetry, and to the analysis of the heat capacity. Neither of the studies reported the computation of the intensities in the infrared and/or Raman spectra of the system, and the refinements of

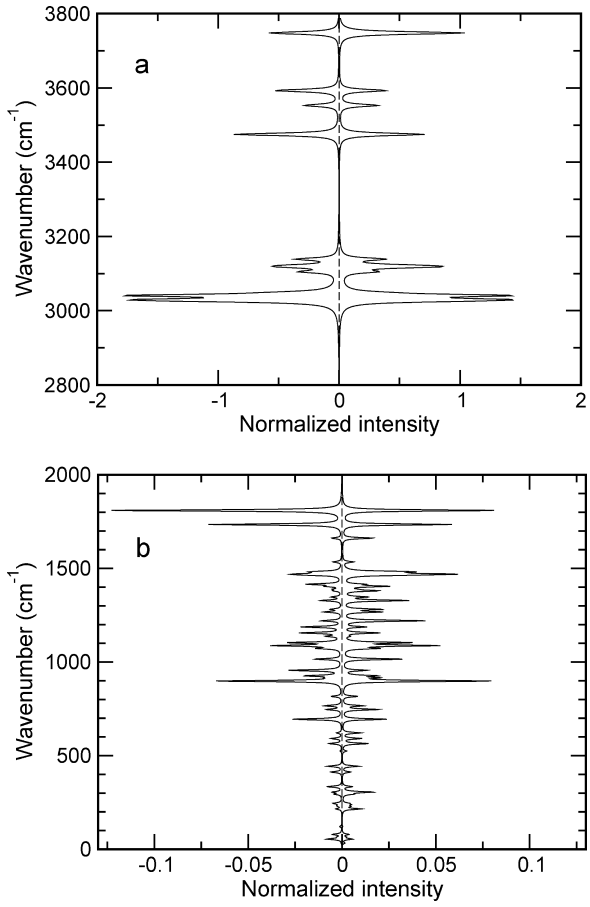


Figure 7. High-frequency (a) and low-frequency (b) parts of the Raman activity spectrum of dialanine molecule calculated with DFT (positive values of *x*-axis) and with BPM (negative values of *x*-axis). The intensity in the spectra were normalized on the total intensity and multiplied by 100; the intensities calculated with the BPM were made negative for comparison.

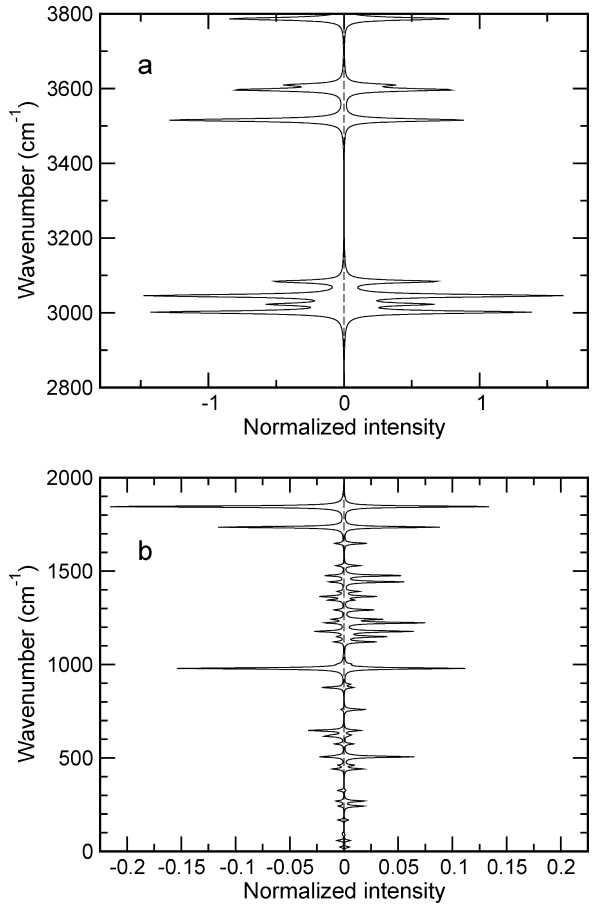


Figure 8. High-frequency (a) and low-frequency (b) parts of the Raman activity spectrum of diglycine molecule calculated with DFT (positive values of *x*-axis) and with BPM (negative values of *x*-axis). The intensity in the spectra were normalized on the total intensity and multiplied by 100; the intensities calculated with the BPM were made negative for comparison.

the force field parameters were therefore carried out solely on the basis of the comparison of vibrational frequencies. In order to evaluate the influence of the force field on the calculated intensities two force fields of a different nature were chosen: a valence force field by Dwivedi and Krimm[33] (100 internal coordinates, 70 force constants) and a Urey−Bradley force field (148 internal coordinates, 31 force constants) used by Porwal et al.[35] The system treated with the former force field was a three-dimensional periodic model of PGI structure, whereas the calculations using the latter force field considered a single infinite polyglycine chain. In both models the coordinates of atoms corresponded to the low molecular weight PGI structure determined by Kajava.[36]

The inspection of the calculated frequencies indicates a good reproduction of the data cited in the previous works and a good agreement with experimentally determined frequencies. However, the potential energy distribution significantly varies between the force fields although the frequencies and general form of the modes are very similar. This fact finds confirmation in Figure 10, which presents the computed Raman spectra in the region 2000−200 cm^{-1} and compares them with the experimental spectrum adapted from ref 37. All the spectra reveal the most intense Raman peaks at ca. 1000 cm^{-1}. In the spectral range from 2000 to 1000 cm^{-1} the relative intensities of the Amide I and Amide II peaks are well reproduced by the calculations with the former being more intense, whereas the intensity of the Amide III peak is underestimated. The weakness of the intensity model discussed above results in an underestimation of the Raman intensity of the peaks due to the angle deformation modes involving CH bonds. Below 1000 cm^{-1} an intense peak is present at 700−730 cm^{-1} in the calculated spectra, while it is absent in the experimental one; the relative intensities of other peaks vary but are in a fair agreement.

Analysis of the potential energy distribution in the 700 cm^{-1} mode shows that it involves the C−Cα stretching and C═O in-plane bending vibrations coupled with angle-bending deformations Cα NC + CCα N of the chain. 90% of the Raman activity comes from the anisotropy of the C═O bond polarizability tensor, and thus the Raman activity of the mode is almost entirely due to the in-plane bending vibration of the bond. Table 1 shows that the EOPs determining the anisotropy of C═O bond polarizability have relatively low statistical uncertainties and that values of the anisotropy obtained in the present work and experimentally determined well agree with each other. Consequently, the presence of the peak with such a large intensity in the

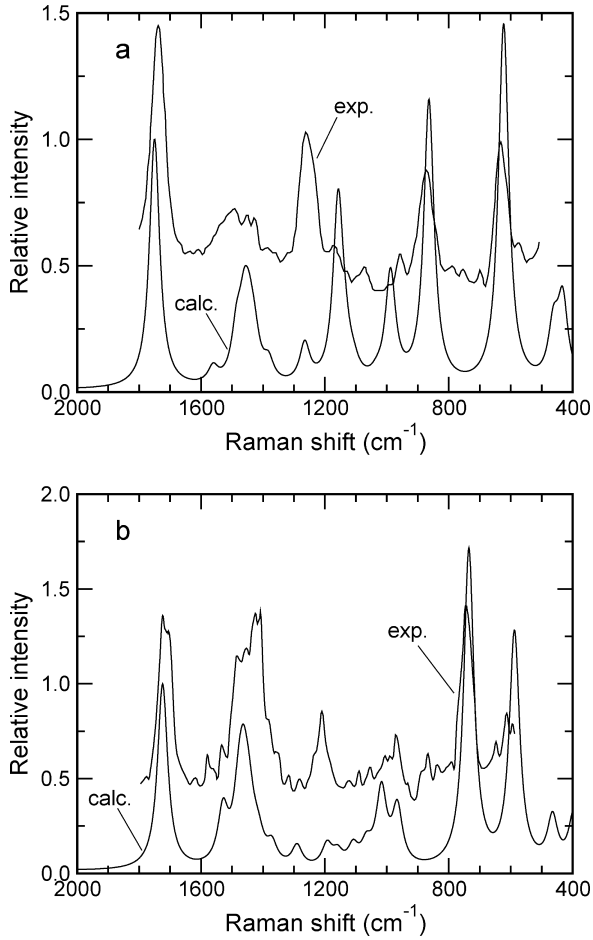


Figure 9. Experimental[28] and calculated Raman scattering spectra of N-methylacetamide (a) and N,N-dimethylacetamide (b) molecules.

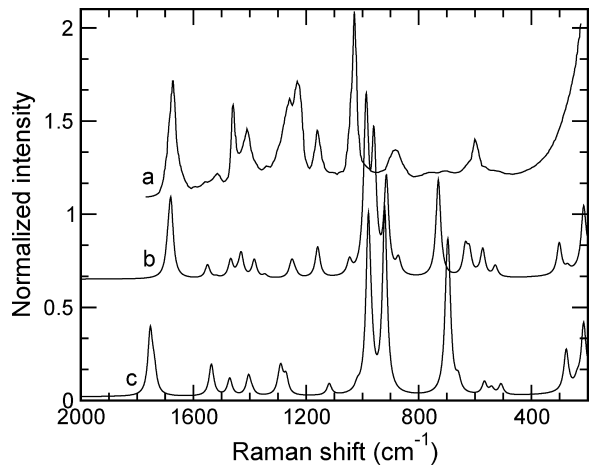


Figure 10. Experimental (a) and calculated Raman spectra of polyglycine I: (b) spectrum computed with a valence force field, ref 33; (c) spectrum obtained with a Urey−Bradley force field, ref 35. The experimental spectrum was adapted from Figure 1 of ref 37. The spectra were shifted along the *y* axis for clarity.

spectrum can unlikely be due to incorrect values of the electro-optical parameters. To our mind, this fact rather points to a problem in the assignment of the experimental bands and to the necessity of performing a new refinement of polyglycine force fields based on an alternative assignment. The corresponding work should go along the following lines:

• precise definition of the structure of experimental samples (e.g., polyglycine I or II) and recording of the vibrational spectra with the use of isotopic substitution permitting an increase in the number of observables,

• use of quantum-chemical calculations to identify the most relevant force constants having a physical sense; the step should also include a precise definition of the set of internal coordinates used in the force field development,

• refinement of the force constants using the experimental data and the parametric models for the spectral intensities. This stage should be followed by an ultime test of the force field transferability.

Such a complete study is well out of the scope of the present work dealing with the EOPs for polypeptides; an example of such a development for aluminosilicate structures can be found in refs 38 and 39. The present results clearly indicate that the computation of the Raman spectra can be of significant help for judging the force field quality.

Besides the limitations inherent in the bond polarizability model, there is another factor affecting the quality of the calculated Raman activities. As it follows from eq 8, the derivative of the polarizability tensor in a normal mode depends on the $(\partial s_j/\partial Q_k)$ quantities that are, in their turn, defined via the derivatives of the Cartesian coordinates with respect to Q_k (and *vice versa*). Obviously, these quantities depend on the potential model used in the vibrational analysis, and thus the choice of the potential model can have a consequence on the computed Raman activities, as it can be seen in Figure 10. To further assess such an influence we have performed a calculation of the Raman activities of the modes of NMA molecule using the vectors of Cartesian displacements obtained in the vibrational analysis of the molecule at the MP2/6-311G++(3df,3pd) level. The comparison of the activities with those computed with the B3LYP/Sadlej basis set level shows that for some modes the R_k values can vary up to 15−20%, and therefore the choice of the force field plays an important role for the Raman intensities computed with the BPM model. The 15% confidence interval is likely the precision that can be achieved with parametric models for Raman intensity calculations.

4. Conclusions

A procedure for obtaining electro-optical parameters of bond polarizability model from results of quantum-chemical computations was applied to derive the EOPs for chemical bonds relevant to biologically important molecules, particularly to peptides. Calculation of the mean polarizability of four test molecules (N-methylacetamide, N,N-dimethylacetamide, dialanine, and diglycine) has shown that the set of parameters reproduces the reference data with a mean absolute error of 0.116 Å^3 and the maximum relative error of 1.6%. The calculation of the nonresonant Raman scattering activities of the normal modes of the molecules has revealed a good correspondence of the BPM computed activities with the reference ones in the region of C−H and N−H stretching vibrations (2800−3800 cm^{-1}), while a systematic underestimation of the Raman activity was found in the energy range

from 1200 to 1500 cm^{-1} that is characteristic to vibrational modes involving variations of angle-bending internal coordinates. The agreement in this spectral region can be improved by changing the equilibrium EOPs for the bonds involved in the coordinates, however, at the expense of the quality in the polarizability tensor components and in the Raman activities of some other modes. An extension of the model to a first-order BPM is expected to improve the quality of the Raman intensity calculations; work in this direction is in progress. Despite the existing deficiencies the resulting set of electro-optical parameters is capable of describing both the polarizability tensor and its derivatives with respect to variations of geometry of large molecules. This conclusion is confirmed by the comparison of the calculated Raman scattering spectra of the N-methylacetamide and N,N-dimethylacetamide molecules with the experimental ones. The results of the calculation of the Raman spectrum of the polyglycine I structure suggest the necessity of revising the force fields for the system and illustrate the usefulness of the Raman spectra computations in the development of force fields of spectroscopic quality. The EOPs can be used in studies of the conformational dependence of the polarizability of peptides and in the Raman spectra simulations by molecular mechanics or molecular dynamics methods.

The comparison of EOPs derived in the present work with those found in the literature shows sometimes significant differences that can reflect the well-known ambiguity of the parameters obtained from experimental data because of ill-conditioning of the corresponding mathematical problem. Hence, the proposed procedure allows one to choose among mathematically equivalent sets of EOPs that which has a physical sense. Furthermore, the approach permits an easy extension to new types of chemical bonds. Such an extension necessitates additional quantum-chemical computations of molecular models having chemical bonds of the new types that are followed by constrained or unconstrained LLSQ fit.

Acknowledgment. The financial support from the Indo-French Centre for the Promotion of the Advanced Research (IFCPAR) under grant no. 3305-1 is gratefully acknowledged.

Supporting Information Available: Structures of molecules in the training and validation sets optimized at B3LYP/pVTZ Sadlej basis set level. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schweitzer-Stenner, R. *J. Raman Spectrosc.* **2001**, *32*, 711–732.

(2) Schweitzer-Stenner, R. *Vib. Spectrosc.* **2006**, *42*, 98–117.

(3) To, T. T. Molecular Dynamics Study of the Vibrational Spectra of Silica Glasses. Ph.D. Thesis, University of Sciences and Technologies of Lille, Villeneuve d'Ascq, France, 2007. Available online: http://tel.archives-ouvertes.fr/tel-00282362/fr/.

(4) Wolkenstein, M. W. *C. R. Acad. Sci. URSS* **1941**, *30*, 791–794.

(5) Applequist, J.; Carl, J. R.; Fung, K. K. *J. Am. Chem. Soc.* **1972**, *94*, 2952–2960.

(6) Maple, J. R.; Ewing, C. S. *J. Chem. Phys.* **2001**, *115*, 4981–4988.

(7) Wang, J.; Vie, X. Q.; Hou, T.; Xu, X. *J. Phys. Chem. A* **2007**, *111*, 4443–4448.

(8) Truchon, J. F.; Nicholls, A.; Iftimie, R. I.; Roux, B.; Bayly, C. I. *J. Chem. Theory Comput.* **2008**, *4*, 1480–1493.

(9) Gussoni, M. In *Advances in Infrared and Raman Spectroscopy*; Clark, R. J. H., Hester, R. E., Eds.; Heyden & Son: London, 1980; Vol. 6, Chapter 2, pp 61−126.

(10) Smirnov, K. S.; Bougeard, D. *J. Raman Spectrosc.* **2006**, *37*, 100–107.

(11) Smirnov, K. S.; Bougeard, D.; Tandon, P. *J. Phys. Chem. A* **2006**, *110*, 4516–4523.

(12) Long, D. A. *Proc. Roy. Soc. (London)* **1953**, *A217*, 203–221.

(13) Volkenstein, M. V.; Gribov, L. A.; Eliashevich, M. A.; Stepanov, B. I. *Molecular Vibrations*; Nauka: Moscow, 1972.

(14) Long, D. A. *The Raman effect*; John Wiley & Sons: Chichester, 2002.

(15) Gough, K. M.; Dwyer, J. R. *J. Phys. Chem. A* **1998**, *102*, 2723–2731.

(16) Gough, K. M.; Dwyer, J. R.; Dawes, R. *Can. J. Chem.* **2000**, *78*, 1035–1043.

(17) Sadlej, A. J. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995–2016.

(18) Sadlej, A. J. *Theor. Chim. Acta* **1992**, *79*, 123–140.

(19) Halls, M. D.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 8819–8824.

(20) 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 D.01*; Gaussian, Inc.: Wallingford, CT, 2004.

(21) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: New York, 1992; pp 670−675.

(22) Applequist, J.; Quicksall, C. O. *J. Chem. Phys.* **1977**, *66*, 3455–3459.

(23) Van Hemert, M. C.; Blom, C. E. *Mol. Phys.* **1981**, *43*, 229–250.

(24) Furer, V. L.; Alekseev, V. V. *Zh. Prikl. Spektrosk. (USSR)* **1986**, *45*, 951–955.

(25) Yokoyama, I.; Miwa, Y.; Machida, K. *J. Phys. Chem.* **1991**, *95*, 9740–9746.

(26) Yokoyama, I.; Miwa, Y.; Machida, K.; Umemura, J.; Hayashi, S. *Bull. Chem. Soc. Jpn.* **1993**, *66*, 400–413.

(27) Montero, S.; Bermejo, D. *Mol. Phys.* **1976**, *32*, 1229–1232.

(28) Triggs, N. E.; Valentini, J. J. *J. Phys. Chem.* **1992**, *96*, 6922–6931.

(29) Fukushima, K.; Ideguchi, Y.; Myazawa, T. *Bull. Chem. Soc. Jpn.* **1963**, *36*, 1301–1307.

(30) Gupta, V. D.; Trevino, S.; Boutin, H. *J. Chem. Phys.* **1968**, *48*, 3008–3015.

(31) Abbe, Y.; Krimm, S. *Biopolymers* **1972**, *11*, 1817–1839.

(32) Moore, W. H.; Krimm, S. *Biopolymers* **1976**, *15*, 2439–2464.

(33) Dwivedi, A. M.; Krimm, S. *Macromolecules* **1982**, *15*, 177–185.

(34) Fillaux, F.; Fontaine, J. P.; Baron, M. H.; Leygue, N.; Kearley, G. J.; Tomkinson, J. *Biophys. Chem.* **1994**, *53*, 155–168.

(35) Porwal, V.; Misra, R. M.; Tandon, P.; Gupta, V. D. *Indian J. Biochem. Biophys.* **2004**, *41*, 34–39.

(36) Kajava, A. V. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *D55*, 436–442.

(37) Small, E. W.; Fanconi, B.; Peticolas, W. L. *J. Chem. Phys.* **1970**, *52*, 4369–4379.

(38) Ermoshin, V. A.; Smirnov, K. S.; Bougeard, D. *Chem. Phys.* **1996**, *209*, 41–51.

(39) Creton, B.; Bougeard, D.; Smirnov, K. S.; Guilment, J.; Poncelet, O. *J. Phys. Chem. C* **2008**, *112*, 10013–10020.

CT800510Y

JCTC Journal of Chemical Theory and Computation

Computer Simulations of Peptides from the p53 DNA Binding Domain

Mey Khalili[†,‡] and David J. Wales*[,§]

Center for Cancer Research Nanobiology Program, National Cancer Institute, Frederick, Maryland 21702, MITRE Corporation, 7515 Colshire Drive, McLean, Virginia 22102, and Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Received December 5, 2008

Abstract: We have studied the dynamics and thermodynamics of two of the four evolutionarily conserved segments from the p53 DNA binding domain, using molecular dynamics and replica exchange simulations. These two regions contain well-defined elements of secondary structure (a β hairpin for region II and an α helix for region V) and bind to DNA in the intact protein. They are also mutational hot spots. The goal of our study was to determine the stability and folding propensity of these peptides in isolation. We used three force fields and solvent models (CHARMM19 with EEF1, CHARMM27 with GBMV, GROMOS96 with SPC). The predicted stability, folding propensity, and secondary structures depend upon the potential. Secondary structure predictors identify helical propensity for region II, in agreement with one of the force fields (CHARMM/GBMV). However, the other two potentials favor β structure for this peptide, although the conformations may differ from the crystal. For region V secondary structure predictions are unclear. Only one force field (CHARMM/GBMV) produces low-lying free energy minima that retain some of the α helical structure from the crystal structure. The other two potentials appear to favor β structure for this peptide.

1. Introduction

p53 has been one of the most extensively studied proteins since its discovery in 1977.[1−5] The p53 gene is mutated in a large number of human cancers.[6−9] Furthermore, in many tumor cells that have the wild-type p53 protein, activity is hindered by the overexpression of regulators or viral oncogenes.[10−17] The active p53 protein is a noncovalent tetramer of four 393 residue monomers,[18−20] which serves as a transcription factor involved in regulating the cell cycle, repairing damaged DNA, and initiating cell death through apoptosis.[8] Each monomer contains two folded domains, namely a core DNA-binding domain (DBD, around residues 94−297)[18,21] and a tetramerization domain (around residues 323−360).[20,22] The remainder of the sequence appears to be intrinsically disordered,[8,9,20] including the N-terminal transactivation domain (NTD, residues 1−67)[20,23−25] and the C-terminal negative regulatory domain (residues 360−393).[22] This intrinsic disorder may reflect the multifunctional nature of p53 and its "promiscuous" binding to a variety of other proteins.[9,26] In the present work our focus is on two peptides from the DBD, which binds to specific elements of the p53 gene promoters.[27,28] This domain has been crystallized,[18,29,30] and its NMR structure has been determined with and without[26] its consensus DNA sequence.

The tumor suppressor activity of p53 arises from its ability to act as a transcription factor and induce the expression of a number of proteins that are involved in DNA repair or the inhibition of cell proliferation and apoptosis.[4,8,31] In normal cells p53 is closely regulated and has a high turnover (half-life between 5 and 40 min),[32] due to the inhibitory negative feedback mechanism of the oncoprotein mdm2 (murine double minute clone 2, human equivalent hdm2).[33] mdm2 binds to p53 in the NTD and thereby inhibits transcriptional

* Corresponding author e-mail: dw34@cam.ac.uk.
[†] National Cancer Institute.
[‡] MITRE Corporation.
[§] Department of Chemistry.

10.1021/ct8005387 CCC: $40.75
Published on Web 04/10/2009

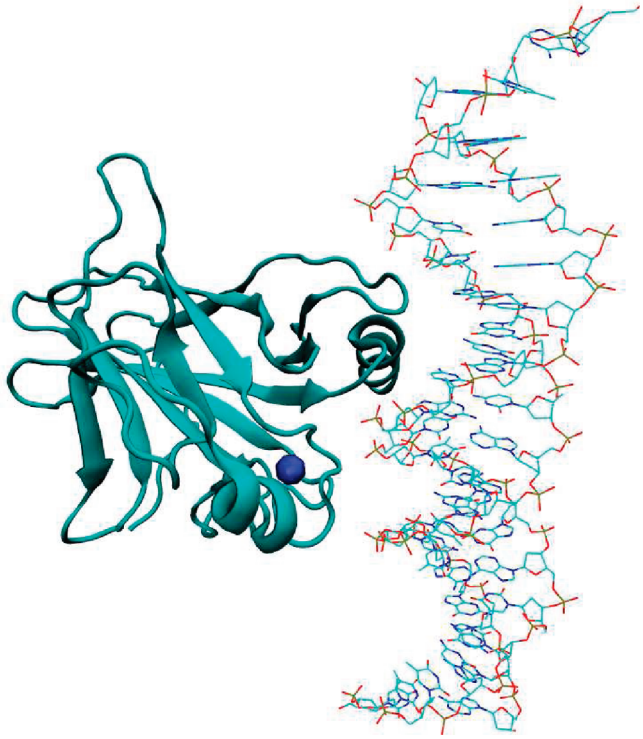


Figure 1. Monomer of the p53 DNA binding domain bound to its consensus DNA-binding site. The zinc cation is shown in dark blue.

activity;[34,35] a second binding site has also been identified for hdm2 in the DBD.[36] High levels of mdm2 may promote polyubiquitination of p53, marking the protein for degradation through the proteasome pathway.[37,38]

Given the crucial role that p53 plays in cell cycle regulation, and the large number of cancers associated with the malfunction of this protein, it is not surprising that there are many ongoing research efforts to rescue or mimic its function in cancer cells.[8,9,17,39] In the most direct application of gene therapy, tumor suppressor genes are expressed in the cells where these genes are defective. For TP53 (the gene encoding the p53 protein), encouraging results have been obtained for some classes of tumor.[40,41] Gene therapy using adenovirus mediated wild-type p53 gene transfer has achieved some success.[42−45] However, this approach does not work for tumors in which the p53 protein is functional, but its activity is hindered by the overexpression of its regulators. For example, some cancers are caused by excess levels of mdm2 or hdm2.[46,47] For these classes of tumor, research has focused on small peptides that bind to the NTD and prevent hdm2 from binding to p53.[17,48]

The crystal structure of the DNA-binding domain of human p53 (PDB code: 1TUP) was determined in a complex with consensus DNA by X-ray crystallography in 1994.[18] It consists of an immunoglobulin-like β sandwich composed of two antiparallel β sheets packed face-to-face across a hydrophobic core[18] (Figure 1). This domain contains four of the five regions that are conserved across species[18,49] (Figure 2), which are numbered II−V, following Cho et al.[18] Regions III (residues 171−181), IV (residues 234−255), and V (residues 270−289) bind to DNA directly. Region II (residues 117−142) binds to DNA through a small segment

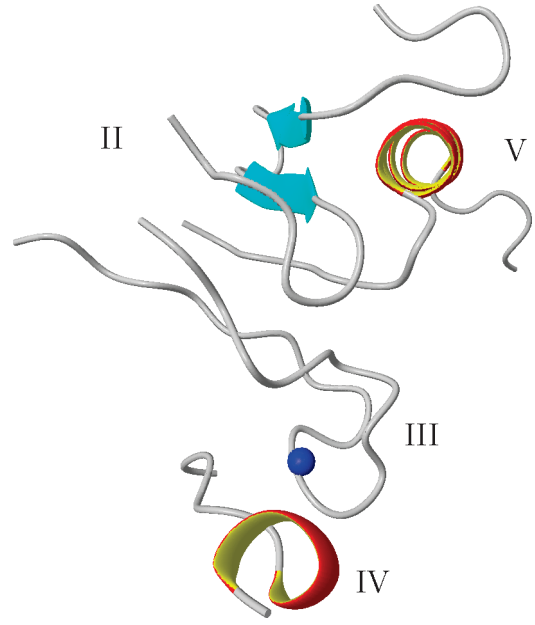


Figure 2. Conserved sites in the DNA-binding domain of p53.[18] Together these sites constitute four of the five conserved regions in a p53 monomer.

(residues 117−123) and corresponds to a loop and hairpin denoted as L1 (residues 112−124) and S2−S2′ (residues 124−142).[18] Region V corresponds to the end of a strand denoted as S10 (residues 271−274) and the α helix H2 (residues 278−286 in the 1994 crystal structure).[18] Regions II and V therefore contain well-defined secondary structures, and form a characteristic loop-sheet-helix motif,[18] where conserved residues make specific contacts with the major groove of DNA.[9] The DBD binds specifically to a palindromic double-stranded DNA promoter site, containing two decameric motifs of the form 5′-PuPuPuC(A/T)(A/T)GPyPyPy-3′ separated by up to 13 base pairs, where Pu is A/G and Py is T/C.[27]

Around 50% of human cancers involve mutations of the p53 gene.[8,9,39] Ninety-five percent of these correspond to the DBD, and 75% are single missence mutations.[39] Examples include[39] R175H (in region III, which disrupts the zinc binding domain), C242S (replaces a zinc ligand in IV), G245S (IV), R249S (IV), M237I (IV), R273H (V), and R282W (V). Arg-248, which makes a minor groove contact from a loop denoted L3, is another mutational hot spot.[18] In general, p53 mutations can be divided into three categories:[9,50] (i) DNA-contact mutations that have little effect on the folding and stability of the DBD (e.g., R273H); (ii) mutations that cause local distortions, mainly in the proximity of the DNA-binding site (e.g., R249S, destabilization $\sim$ 2 kcal/mol); (iii) mutations that cause global unfolding (in the β sandwich), which are destabilized by more than 3 kcal/mol. Bullock and Fersht note that mutations in the DNA-contact regions are oncogenic in any of these three categories, while tumorigenic mutations in the β sandwich fall into category iii.[39] For example, zinc-binding-site mutations in the DBD may have a distorted structure that fails to bind DNA.[39,50] One current research goal is to find small molecules that can rescue the function of destabilized p53 mutants by a mass action effect, namely binding strongly to the native state,

but only weakly to the denatured protein.[8,9,39] This strategy will be particularly challenging for DNA-contact mutants and mutations that perturb the zinc-binding region.[9] Alternatively, a superstable quadruple mutant of the p53 DBD has been designed, where an increase in structural rigidity results in increased thermostability and restoration of activity in several oncogenic mutants.[51,52] This superstable mutant has been used to demonstrate that hot-spot mutations have the same effect upon stability in the isolated core domain and full-length protein.[53]

During the present study, the structure of the free p53 core domain solved with NMR spectroscopy was published by Fersht and co-workers,[21] who characterize it as a marginally stable protein at body temperature, making p53 particularly susceptible to inactivation by destabilizing mutations.[9] Compared to the crystal structure solved in 1994, the α helix that belongs to region V is more than one turn longer in the NMR structure. Secondary structure predictors indicate that this longer version of region V might form a helix in isolation (Table 3). The calculations for region V in the present work all refer to residues 272−289 identified in the 1994 crystal structure.[18] Future work will consider a region V peptide that includes all the residues characterized as helical in the latest NMR structure.[21]

Our aim in this paper is to study the foldability and stability of peptides corresponding to regions II and V as isolated fragments. In native p53 the H2 α helix and the S2−S2′ β hairpin are packed together, and our motivation is partly to see whether this packing is required to stabilize the two elements of secondary structure.

It is well known that the force fields used to study biological molecules are not perfect. There have been a number of studies linking the choice of force field to differences in structural propensities in biological molecules. For example, Eisenmenger and Hansmann performed thermodynamic studies of Met-enkephalin and found subtle differences in the energy landscapes obtained for the ECEPP/2 and ECEPP/3 force fields.[54] Lwin and Luo studied a β hairpin with several alternative AMBER potentials, and with implicit and explicit solvents. They found significant differences between the implicit and explicit solvent simulations.[55] Yoda et al.[56] examined the secondary structure propensities of six protein force fields and also found clear differences. The alternative structures and energy landscapes that we have characterized for the two p53 peptides are discussed in the Results section.

2. Methods

2.1. Explicit Solvent MD Simulations. Explicit solvent simulations were performed with the GROMACS software.[57,58] We used the GROMOS96 53a6[59] force field, which has been parametrized to reproduce the enthalpies of hydration and polar solvation for a range of compounds. Each fragment was solvated by adding a cubic box of 216 SPC flexible water molecules[60] up to 9 nm from the periphery of the molecule. The volume of the box was chosen so that the density was approximately 100 g/L. Chloride and sodium ions were added to neutralize each system, and an initial minimization was performed by constraining the fragment and running molecular dynamics for the water molecules. Each system was then simulated with periodic boundary conditions. The temperature was held constant by the Berendsen thermostat,[61] and the Lincs algorithm[62] was used to constrain the hydrogens. The particle mesh Ewald method was employed[63,64] for the electrostatics and the van der Waals terms with a cutoff of 14 Å. The cutoff distance for the short-range neighbor list was set to 9 Å. The cutoff distance for the Lennard-Jones short-range neighbor list was set to 14 Å. Snapshots were collected every 500 ps with an integration time step of 2 fs. The neighbor list was updated every five steps. For the high temperature simulations at 340 K, the coupling constant to the Berendsen barostat was increased from 0.5 to 0.9 to allow for larger fluctuations in the pressure.

2.2. Implicit Solvent MD Simulations. Implicit solvent simulations were performed with the CHARMM package.[65,66] To provide additional comparisons, we considered two different implicit solvent models, namely EEF1[67] and GB-MV.[68] Implicit water solvation with EEF1 is incorporated into the CHARMM19 polar hydrogen function.[69] In this force field, the effective energy for a given conformation of the protein is the free energy of the system consisting of the macromolecule and the solvent averaged over all the solvent degrees of freedom at a given temperature. The total free energy of the protein-solvent system is the sum of the average effective energy and the configuration entropy. The ability of EEF1 to represent solvent effects was demonstrated by Lazaridis and Karplus, who were able to reproduce conformations for proteins that are comparable to those in explicit solvent simulations.[67] EEF1 has also been able to produce results similar to those of explicit solvent simulations in folding/unfolding studies.[70] Furthermore, the electrostatic energy obtained with EEF1 correlates very well with the screened Coulomb potential-implicit solvent model and reasonably well with the Poisson−Boltzmann energies.[71]

To increase the efficiency of the molecular dynamics (MD) simulations with the EEF1 solvation model, we used the multiple time step (MTS) algorithm with Langevin dynamics.[72] In this method, which is based on Trotter factorization, force linearization is combined with force splitting techniques. Unlike the conventional multiple time step algorithm,[73] this method merges the slow and fast motions via extrapolation rather than impulses, which results in a more significant time step increase. The combination of Langevin dynamics and the EEF1 implicit solvent enabled us to obtain relatively long simulation times. The runs were performed at 300 K with the MTS 4 24 integration scheme. The time steps were 0.5 fs for the fast motions, 2 fs for the intermediate motions, and 48 fs for the slow motions, yielding a 48 fs time step overall. The cutoff for short-range forces was 6.0 Å.

The generalized Born (GB) implicit solvent model is one of the most successful methods for approximating solvation energies. The most important, and time-consuming, part of this theory is the calculation of Born radii, which define the spherically averaged distance of each atom to the solvent boundary. Still and co-workers[74] introduced an expression for the Born energy, which often provides good agreement

Table 1. Secondary Structure Prediction for Region II with Different Servers[a]

	GTAKSVTCTYSPALNKMFCQLAKTCP
Jpred[88]	−−−−−EEE−−−HHHHHHHHH−−−−−−
PSIPRED[89]	CCCCEEEEECCHHHHHHHHHCCCCC
GoreIV[90]	CCCCCEEEECCCCCCCCCCCEEEEEC
HNN[91]	CCCCEEEEECCHHHHHHHHHHHCCC
nnpredict[92]	−−−−−EEE−−−−−−HHHHHHH−−−−−
porter[93]	CCCCCCCEEEECCCCEEEECCCCCEC
SOPMA[94]	TCCHHEEECCCHHHHHHHHHHTCCT
SWISS-MODEL[95]	CCCEEEEEEECCCHHHHHHHHHCCC
SCRATCH[96]	CCCCEEEECCCHHHHHHHHHHCCCC

[a] The top of the table contains the sequence. The first column lists different secondary structure servers used for predictions. The letters in the rows in the second column correspond to the propensity of each residue in the sequence: C, coil; E, extended; H, helical; T, turn. Most servers indicate that region II has helical propensity.

with the Poisson energy. The advantage of the Still equation is that, unlike the Poisson equation, it is an analytic function of atomic positions and can be used in algorithms that require force calculations. For GB simulations we used the CHARMM27 force field, which includes the ϕ and ψ grid correction CMAP surface term.[75] We used the GBMV implicit solvent model[68] with the γ surface tension factor set to 0.015 kcal/mol Å^2 to model the hydrophobic effect.[76] A Nose-Hoover thermostat[77,78] was used to keep the temperature constant. The cutoff for nonbonded list generation was set to 20 Å, the cutoff for nonbonded interactions was set to 18 Å, and the onset of switching for nonbonded interactions was set to 16 Å.

2.3. Replica Exchange Molecular Dynamics Simulations. The thermodynamic properties of the fragments were studied using replica exchange molecular dynamics (REMD).[79−81] Replica exchange simulations were performed with the multiscale modeling tools for structural biology (MMTSB)[82] toolkit. We used the same two force fields that were employed for the implicit solvent simulations, namely CHARMM19 with the EEF1 solvation model and CHARMM27 with the GBMV force field. The number of replicas was set to 32 with temperature values selected exponentially between 200 and 900 K, which produces more replicas around the lower bound. We selected the temperature range to permit efficient sampling of the energy landscape, while allowing for more replicas around the temperature range of interest (298 K). We set the number of replicas to 32 to ensure significant overlap between the energy histograms of adjacent replicas.

The length of the MD simulations in each replica cycle was 20 ps, which should lie above the upper bound for the relaxation time of the water molecules.[83] The replica cycles were repeated until convergence was achieved, as judged from the behavior of the heat capacity curve and the free energy surfaces. The PDB structures were energy minimized and equilibrated at 300 K for 1 ns. They were then used as the initial structure for each replica. The averages for different observables, and the free energies, were calculated using the weighted histogram analysis method.[84]

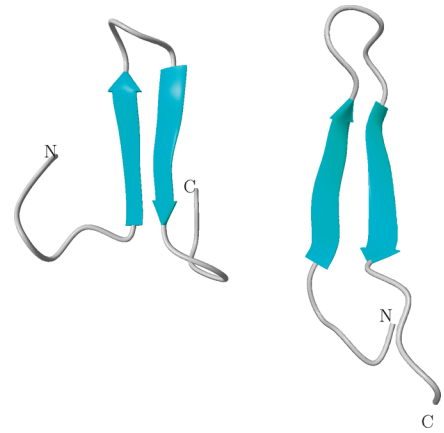


Figure 3. Crystal structure (left) and final structure (right) of region II after 45 ns of MD simulations with CHARMM19/EEF1. At the end of the run the hairpin has elongated by two residues on each side.

3. Results

3.1. Region II: Dynamics. The fastest methods for predicting the secondary structure propensity of a sequence are knowledge-based, homology modeling algorithms.[85] These methods compare a given sequence to a database of sequences for which the structure is known and calculate the statistical propensity of a sequence to form a particular secondary structure. The structure propensity of region II predicted with a number of secondary structure predictors is shown in Table 1. The consensus is that the fragment corresponding to region II is likely to form a helix.

To compare the knowledge-based results with the physics-based methods, we performed MD simulations using CHARMM19/EEF1 and the MTS algorithm at 300 K, to determine whether the hairpin remains folded on a reasonable time scale. The simulation was started from the minimized crystal structure of region II (residues 117−142, Figure 2), and continued for 45 ns.

We chose the hairpin section of region II to calculate the root-mean-square deviation (rmsd), because the rest of the chain is essentially structureless, and fluctuates significantly during the simulation. Examining the rmsd (not shown), it is evident that the hairpin is stable with an rmsd around 2.2 Å during the 45 ns simulation, and does not unfold. Figure 3 (right) shows the conformation at the end of the run, along with the crystal structure (left). During the simulation the hairpin grows from residues 123−136 to span residues 121−139. The results of MD simulations with CHARMM19/EEF1 contrast with the secondary structure predictors (Table 1), which classify region II as helical. The disagreement could be due to the fact that the minimum on the free energy landscape containing the crystal structure is deep enough to prevent the trajectory from escaping at 300 K within 45 ns of simulation. To test this hypothesis, we considered higher temperature MD simulations with CHARMM27/GBMV. We employed an alternative force field at this stage both to provide a comparison with CHARMM19/EEF1 and because the EEF1 parameters are actually fitted specifically for physiological temperatures.[67] As before, these simulations were started from the crystal structure following a local minimization.

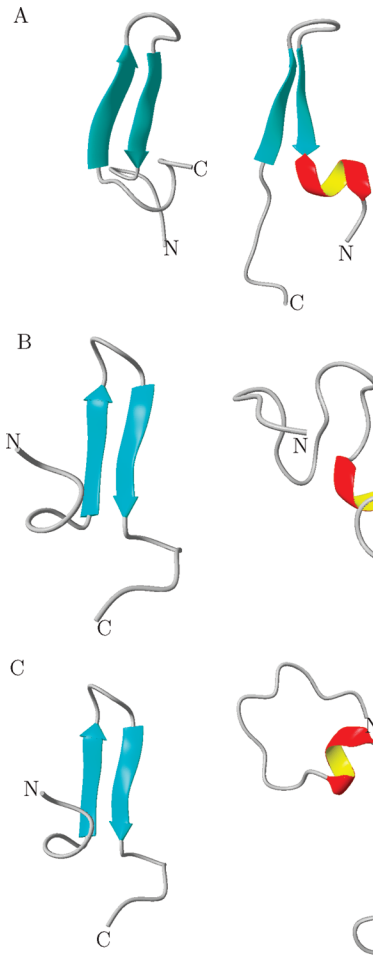


Figure 4. Initial structures (left) and final structures (right) of region II with CHARMM27/GBMV after MD runs at (A) 350, (B) 400, and (C) 450 K. A helical turn appears at 350 K. At 400 and 450 K the hairpin is completely lost, but the helix turn persists, and shifts through the sequence. Since the hairpin is expected to be thermodynamically more stable than a helix,[96,97] this result might indicate some preference for helical conformations.

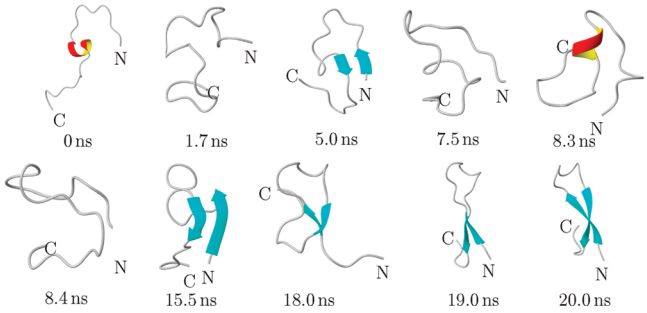


Figure 5. Snapshots of the folding pathway of region II with GROMOS96/SPC at 340 K. There is a clear competition between the formation of a hairpin and a helix during the first 9 ns of the simulation.

With CHARMM27/GBMV, region II is stable on the MD time scale at 350 K (rmsd never increases above 3.5 Å, data not shown), but unfolds at higher temperatures (rmsd higher than 6.0 Å, data not shown). The final structure of region II at the end of the 350 K simulations includes the original hairpin (Figure 4). However, a helical turn appears at the N-terminus. At 400 and 450 K the hairpin is lost in about 1.5 and 1 ns, respectively. However, the helix turn persists, and it shifts its position from the N-terminus to the C-terminus of the unfolded sequence (Figure 4B,C, right). The persistence of the helical turns could reflect a structural preference, or a bias in the force field toward helical conformations. Either way, the behavior of region II with CHARMM27/GBMV is more consistent with the secondary structure predictors, which classify this region as helical (Table 1).

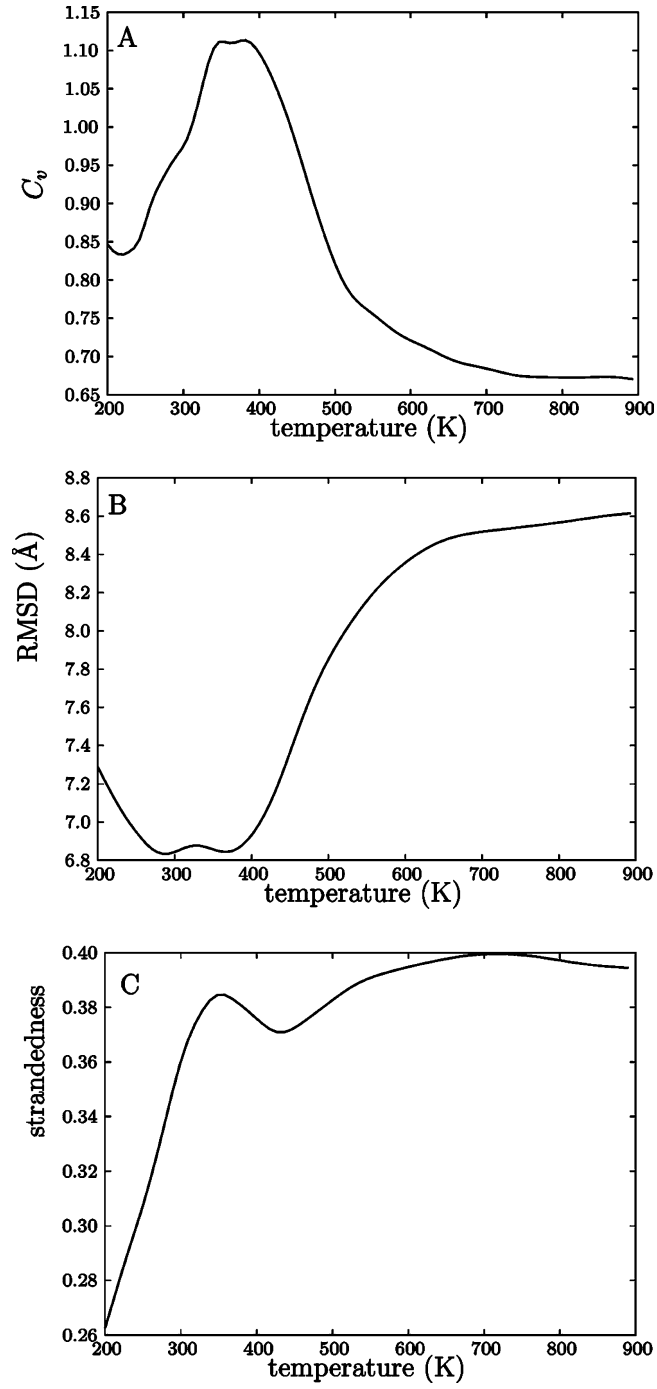


Figure 6. Observables calculated in REMD simulations of region II with CHARMM19/EEF1. (A) Heat capacity [kcal/(mol K)], (B) rmsd from the crystal structure, and (C) strandedness.

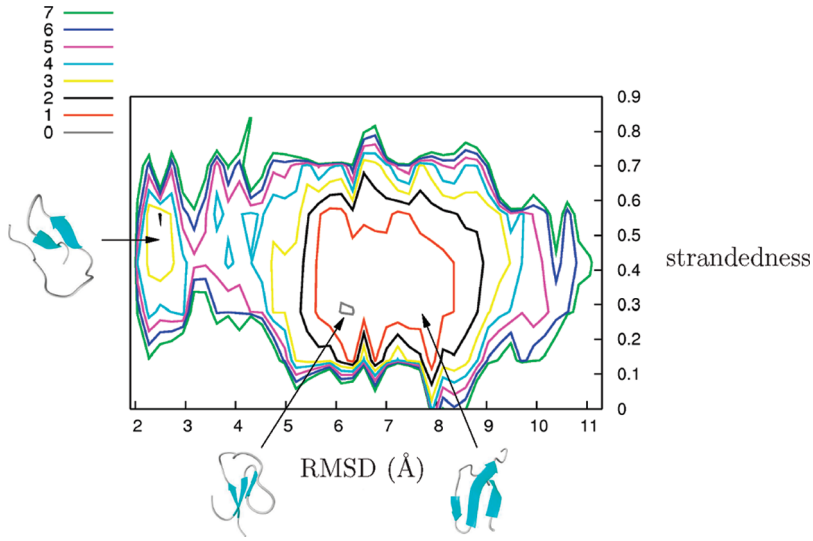


Figure 7. Two-dimensional free energy surface (kcal/mol) of region II as a function of rmsd and strandedness obtained using CHARMM19/EEF1 at 298 K.

Next we tried to fold region II from an unfolded conformation. We used the final structure obtained at 450 K (Figure 4C) as the starting conformation. This choice was made to help the helix nucleate and elongate, if indeed the helical conformation is favorable for this sequence. We used the GROMOS96 force field and explicit SPC water. We also chose a folding temperature of 340 K, to enhance sampling and speed up the folding. We note that this choice might bias the sequence toward β conformations. The simulations were performed for 21 ns. The rmsd of the hairpin from the crystal structure starts around 7.5 Å and drops to about 5.7 Å at about 8 ns, fluctuating around this value thereafter. Figure 5 shows snapshots of region II along its folding pathway. Initially, there is competition between the formation of helix and hairpin. The helical turn disappears at about 2 ns, and a hairpin forms, but does not persist. The helix reappears after 8.3 ns, but disappears again quickly. The same hairpin appears briefly and elongates, but the structure readjusts so that the hairpin corresponding to the crystal structure nucleates at about 18 ns and elongates. The final hairpin is slightly shifted (the turn is shifted by two residues in the C-terminus direction) from the crystal structure. Hence, with the GROMOS96 force field, the sequence appears to favor a hairpin, in contrast to the secondary structure predictions, but in agreement with the crystal structure.

3.2. Region II: Thermodynamics. To study the thermodynamics of region II, we performed REMD simulations with both the CHARMM19/EEF1 and CHARMM27/GBMV force fields. With CHARMM19/EEF1, the system reached equilibrium after 2.36 μs combined time (3685 replica cycles for each of 32 temperatures, each consisting of 10 000 MD steps with 0.002 ps time step). Figure 6 shows different observables at the end of the simulations. The heat capacity curve (Figure 6A) is broad, suggesting the absence of a sharp transition to a single minimum below the folding temperature. Figure 6B shows the average rmsd of the hairpin from the crystal structure at different temperatures. At all temperatures the rmsd is more than 6.8 Å from the crystal conformation. It dips around the physiological temperature range, which indicates that the force field produces more native-like conformations here. Before the system "melts", it adopts conformations that are about 7 Å from the crystal structure. Visual inspection of the low-energy conformations indicated that almost all of them were β structures. We therefore examined a different observable to measure the propensity of the sequence to form extended conformations.

Calculating the extent of β conformations present in MD snapshots is more difficult than for α helices. In particular, changes that are relatively minor upon visual inspection, such as a slight shift in the hairpin formation, can produce large rmsd values. β strands also involve well-separated portions of a chain, making a distance measure more difficult. We used the ϕ and ψ angles to measure the propensity of the chain to form extended structures, following Yoda et al.,[56] who measured these two angles and compared them with the values of the Ramachandran map corresponding to extended structures. We considered a residue to be in the β (or extended) state if the backbone dihedral angles (ϕ and ψ) were in the ranges $-130° \pm 50°$ and $135° \pm 45°$, respectively. We then counted the number of residues with dihedral angles in this range, and divided by the total number of residues in the chain. We call this measure "strandedness", following Yoda et al. Usually a strandedness of 0.4 or higher means that most of the chain is extended, and not surprisingly, strandedness increases with temperature. The strandedness of the crystal conformation is 0.24 (Figure 3, left), while that of the final structure obtained from the MD simulations with CHARMM19/EEF1 is 0.44 (Figure 3, right). Figure 6C shows that the strandedness of region II increases with temperature and reaches a maximum of 0.38, after which it decreases before increasing again as the chain unfolds. This maximum near the physiological temperature is another indication that the force field favors extended conformations around the physiological temperature.

We also measured helicity, i.e., the propensity of the chain to form a helix, using the number of backbone carbonyl, i, and nitrogen, $i + 4$, atoms separated by less than 4 Å. There are 22 possible contacts in a sequence of 26 residues, and

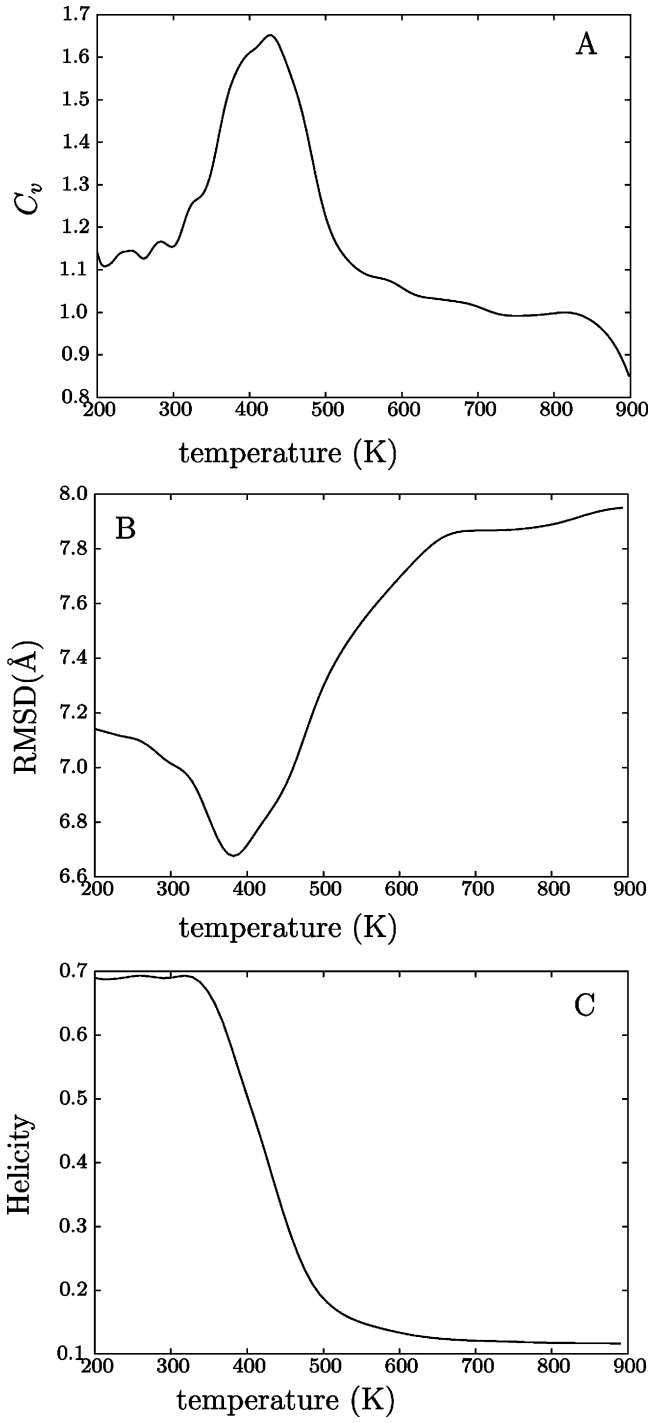


Figure 8. Calculated observables from REMD simulations of region II with CHARMM27/GBMV. (A) Heat capacity [kcal/(mol K)], (B) rmsd from the crystal structure, and (C) helicity. The sequence forms helical conformations with this force field rather than the native β strand.

helicity is defined as the percentage of them that actually form. A plot of average helicity vs temperature (not shown) reveals values below a helicity of 0.045 (corresponding to one helical hydrogen bond) throughout the temperature range considered. Hence the helical content of this sequence is negligible for CHARMM19/EEF1.

Figure 7 shows the two-dimensional free energy surface as a function of rmsd and strandedness at 298 K. The global free energy minimum corresponds to structures that are about

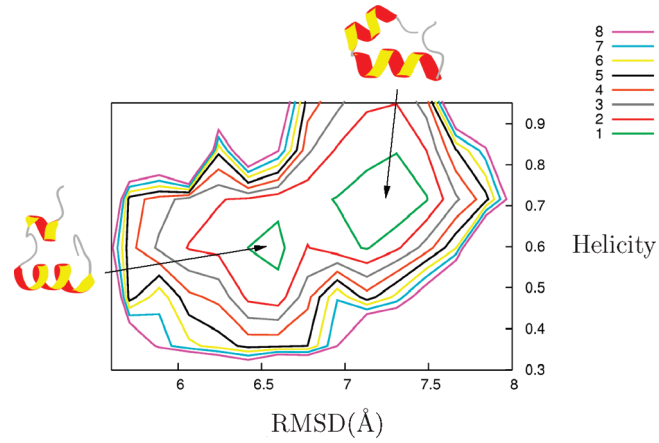


Figure 9. Two-dimensional free energy surface (kcal/mol) for region II with CHARMM27/GBMV at 298 K.

Table 2. Secondary Structure Predictions for Region V (from the 1994 Crystal Structure) with Different Secondary Structure Prediction Servers[a]

	VRVCACPGRDRRTEEENL
Jpred	sequence too short
PSIPRED	sequence too short
GorelV	CCCCCCCCCCCCCEEEEC
HNN	CEEECCCCCCCCCCCCCC
nnpredict	−−−−−HHHHH−−−−−−−−
porter	CEECCCCCCCCCCHHCCC
SOPMA	EEEECCTTCCCCCHHHHH
SWISS-MODEL	CEEEECCCCCCCCCCCCC
SCRATCH	CEEEECCCCCCCCCCCCC

[a] There is no strong consensus on the structure of this sequence.

Table 3. Secondary Structure Predictions for Region V (from the New NMR Structure) with Different Secondary Structure Prediction Servers[a]

	VRVCACPGRDRRTEEENLRKKGEPHH
Jpred	−EE−−−−−−−−−−−−−HHHH−−−−−−
PSIPRED	CCEECCCCCCCCCHHHHHHHCCCCCC
GorelV	CCCCCCCCCCCCHHHHHHHHCCCEEC
HNN	CEEECCCCCCCCCHHHHHHHCCCCCC
nnpredict	−−−−−−−−−−−−−HHHHH−−−−−−−−
porter	CEECCCCCCCCCHHHHHHHHCCCCCC
SOPMA	EEEECCTTCCCCCHHHHHHHTCCCCC
SWISS-MODEL	CEEEECCCCCCCCCHHHCCCCCCCCC
SCRATCH	CEEEECCCCCCCCHHHHHHHCCCCCC

[a] Helical propensity is generally predicted.

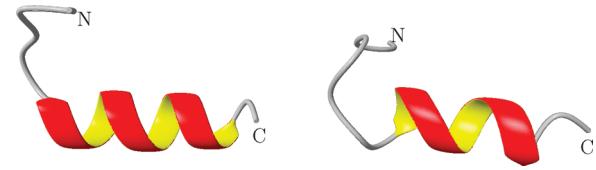


Figure 10. Crystal structure of region V (left) and final structure (right) after 70 ns of MD simulation with CHARMM19/EEF1. The helix loses hydrogen bonds from both ends, while retaining the central part of the helix.

6 Å from the crystal structure and have a strandedness of 0.3. However, this minimum belongs to a broad region of low free energy spanning rmsd values between 6 and 8 Å and strandedness between 0.2 and 0.5. Therefore, at room

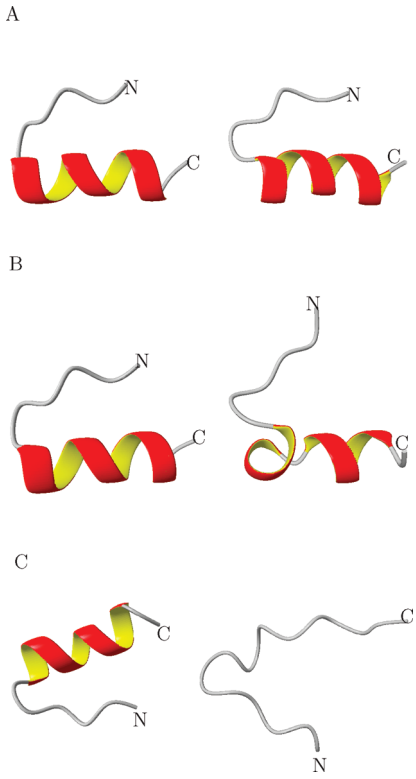


Figure 11. Initial structures (left) and final structures (right) of region V with CHARMM27/GBMV at (A) 350, (B) 400, and (C) 450 K. As for region II, region V remains close to its crystal conformation at 350 K, but changes structure at higher temperatures.

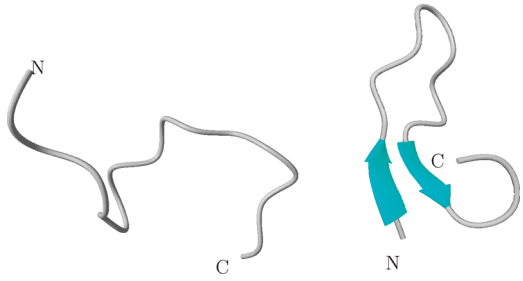


Figure 12. Initial unfolded structure (left) and conformation at the end of MD simulations (right) for region V with the GROMOS96/SPC force field after 20 ns. The sequence adopts a hairpin, which persists to the end of the run.

temperature, the sequence does not have a single well-defined structure. There is also a free energy minimum corresponding to a narrow range of rmsd around 2 Å and a relatively wide range of strandedness. This feature corresponds to the crystal structure, and lies 2 kcal/mol higher than the global minimum, separated by a barrier of about 3 kcal/mol. As expected from the CHARMM19/EEF1 MD simulations, this minimum is deep enough to trap the system at room temperature on short time scales. Although this free energy minimum contains conformations that are comparable to the crystal, it also contains non-native β structures.

To check whether changing the temperature would make the free energy minimum corresponding to the crystal structure more favorable, we recalculated the free energy profile at 310, 390, 390, and 450 K, respectively (data not

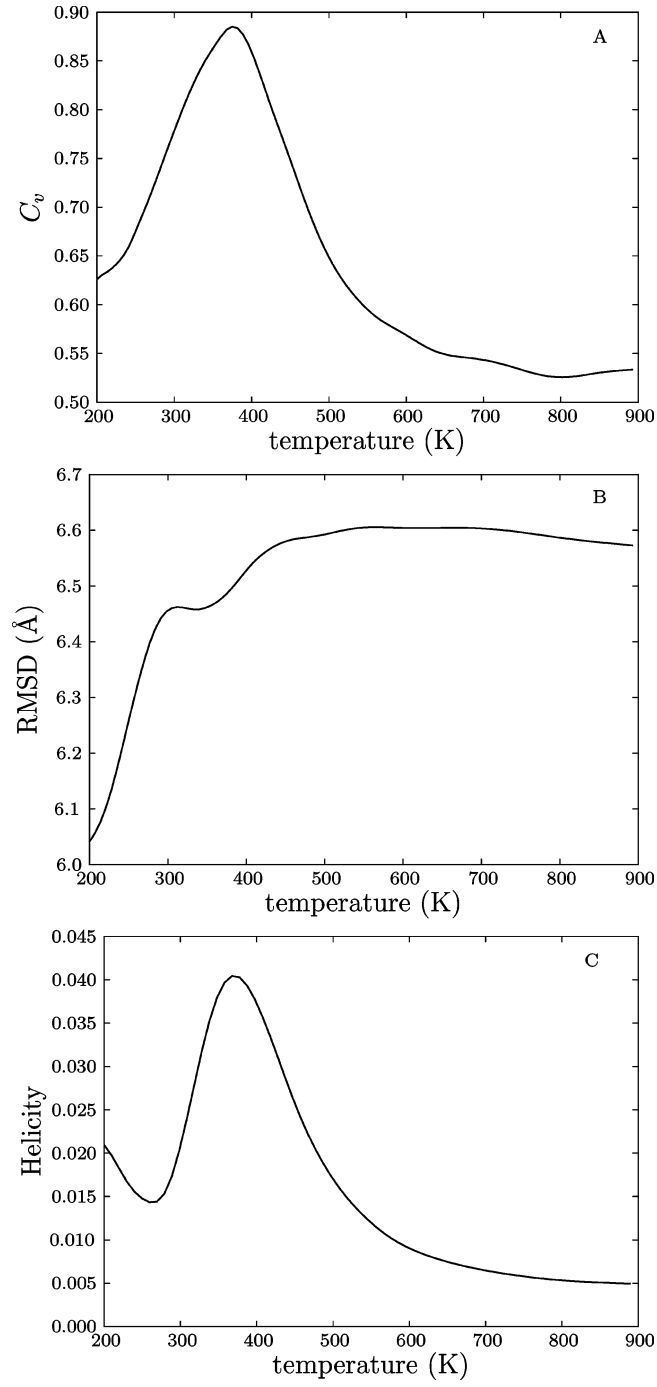


Figure 13. Calculated observables for REMD simulations of region V with CHARMM19/EEF1. (A) Heat capacity [kcal/(mol K)], (B) rmsd, and (C) helicity.

shown). At temperatures above 350 K this minimum becomes progressively shallower, indicating marginal stability. However, the global minimum for 298 K persists and broadens at higher temperature, as expected, indicating that the sequence prefers to be in the more extended conformations at all temperatures with this potential.

Next we repeated the REMD simulations with the CHARMM27/GBMV force field, which gives very different results. The system reached equilibrium after 27.52 ns combined time (43 000 replica cycles for each 32 temperatures, each consisting of 10 000 MD steps with a 0.002 ps time step). Figure 8 shows different observables from this

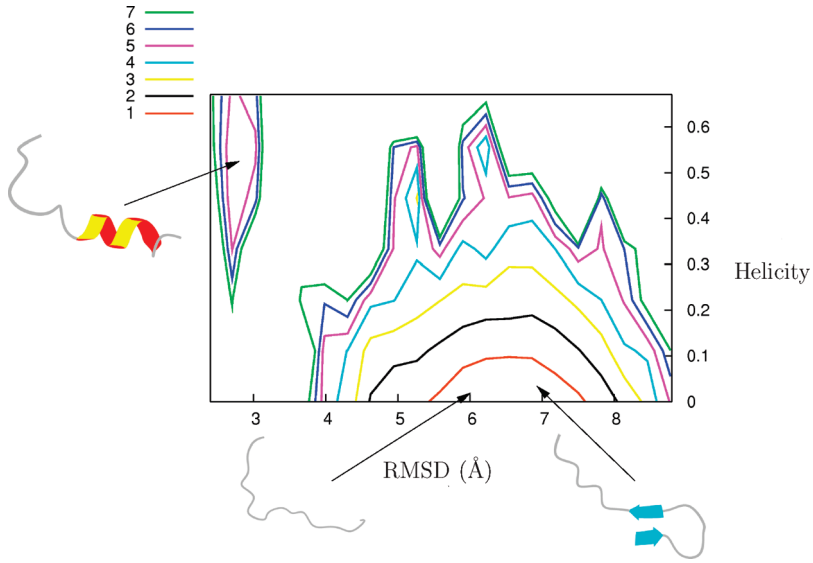


Figure 14. Two-dimensional free energy surface (kcal/mol) of region V for CHARMM19/EEF1 at 298 K, projected onto rmsd and helicity.

simulation. As for Figure 6A, the heat capacity is broad, again indicating no well-defined transition to a single minimum below the transition temperature. At all temperatures the rmsd from the crystal structure (Figure 8B) is more than 6.8 Å, and the average helicity is about 0.7 at temperatures below 400 K, which corresponds to 15 helical hydrogen bonds (Figure 8C). A corresponding two-dimensional free energy surface is shown in Figure 9. There are two minima of equal depth, separated by a small barrier. It is clear that both minima contain conformations with high helicity. It is also clear that they can easily interconvert, since the barrier of 1 kcal/mol is too small to confine the structures in either one of the minima at physiological temperatures. Sequence II in isolation therefore prefers to form helices with this force field, in agreement with secondary structure predictors, and in contrast to the crystal structure and the other two potentials considered. Our MD runs at lower temperatures were clearly not long enough to achieve equilibrium, although the run at 350 K did produce one helical turn, as discussed above. The broad shape of the heat capacity curve and the broad minima on the free energy surface are due to the fact that the N-terminus of the sequence forms many conformations with different amounts of helicity.

3.3. Region V: Dynamics. As mentioned in the Introduction, region V considered here (residues 272−289) is based on the crystal structure obtained in 1994, and does not include the additional helical turn beyond residue 289 reported for the latest NMR structure.[21] Apart from containing many charged residues, which help it to bind directly to DNA, this region is also notable for having a proline nearer the N-terminal part of the sequence (P278). This proline forces the N-terminus to adopt a kinked position relative to the rest of the sequence. In contrast to region II, there is a lack of consensus among secondary structure predictors for residues 272−289 of region V, although most of them predict the sequence to be unstructured in isolation (Table 2). However, when we consider residues 272−297, which include the complete helix from the NMR structure, the homology modeling methods generally predict region V to be helical (Table 3).

We performed 70 ns MD simulations for the 18-residue peptide from region V with CHARMM19/EEF1 using the MTS algorithm. The rmsd fluctuates around 2.5 Å, with relatively minor variations until 20 ns (data not shown). From this point, it increases by 1.5 Å and the fluctuations grow significantly, because the N-terminal and C-terminal parts of the helix unfold somewhat and the ends move more freely. The helicity decreases from unity to about 0.6 in the first 20 ns of the run, and then stabilizes (data not shown). This value corresponds to the loss of approximately three helical hydrogen bonds. Beyond this point the rmsd increases significantly, indicating that first the three helical hydrogen bonds are lost, and then the chain becomes more labile. Figure 10 shows the crystal structure (left), and the structure at the end of the 70 ns MD simulations (right). The hydrogen bonds lost belong to both ends of the helix, while the core remains intact. As mentioned above, due to the position of P278, the orientation of the N-terminus relative to the rest of the peptide rarely changes.

High temperature MD simulations with CHARMM27/GBMV were performed for region V at 350, 400, and 450 K, for intervals of 10, 9, and 9 ns, respectively. As for region II, region V is stable on this time scale at 350 K with the helix intact, but unfolds at higher temperatures. For the 400 K run the helix loses two helical hydrogen bonds by the end of the trajectory; the final structure is a distorted helix with a kink in the middle (Figure 11 B). Helicity is completely lost in the 450 K simulations; the final structure has only one helical hydrogen bond, and lacks any other helical features (Figure 11C).

To see if region V folds to its crystal structure, we considered MD runs using explicit solvent simulations with the GROMOS96 force field and explicit SPC water. We used a temperature of 340 K and ran the simulations for 20 ns to see if we would observe any helix nucleation. The rmsd

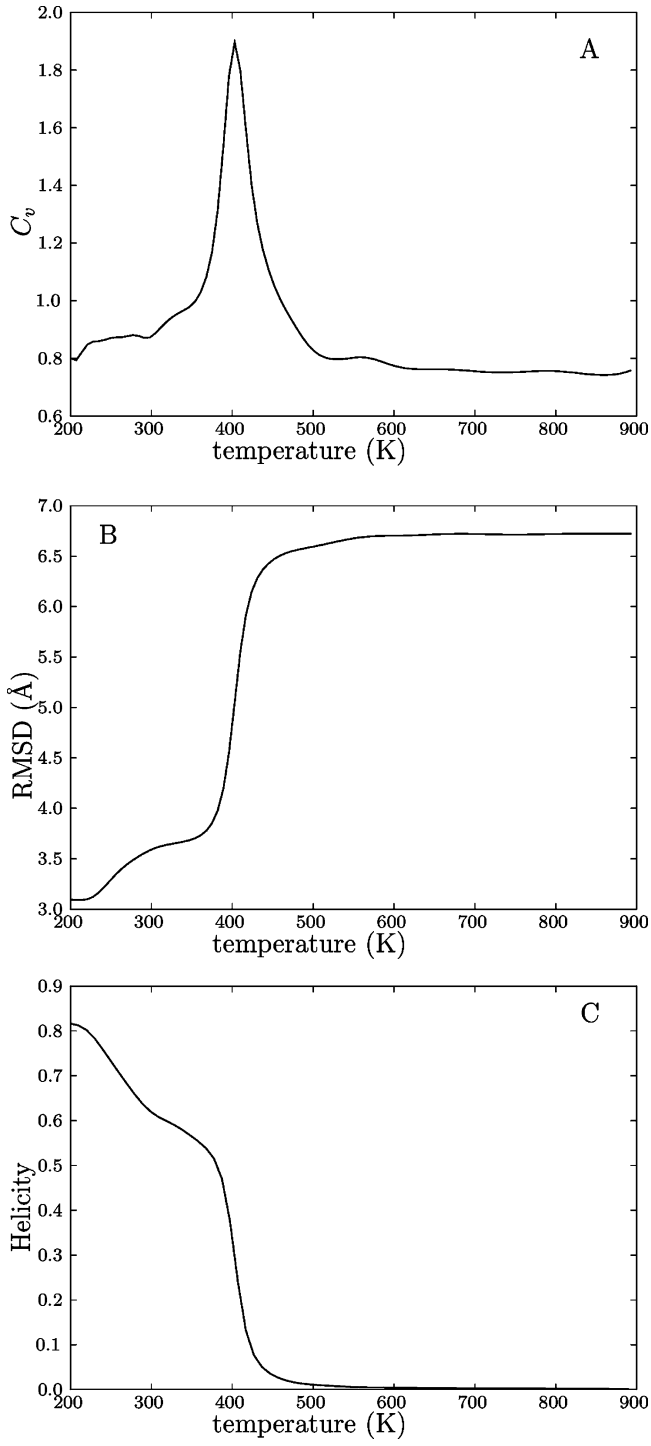


Figure 15. Calculated observables from REMD simulations of region V with CHARMM27/GBMV. (A) Heat capacity [kcal/(mol K)], (B) rmsd of the helical region, and (C) helicity.

during the simulation fluctuates around 6.5 Å, and the helicity does not increase above 0.1 (data not shown), indicating that there is no helical hydrogen-bond formation. Instead we observe a degree of hydrophobic collapse to compact conformations, but because the sequence is highly charged it prefers to maximize its contact with water. This peptide is stable in the hairpin configuration of Figure 12 and remains in this conformation until the end of the simulation for GROMOS96/SPC.

3.4. Region V: Thermodynamics. Figure 13 shows various properties calculated in replica exchange MD simulations for region V with CHARMM19/EEF1. The simulations were performed for 1.5 μs (2417 replica cycles for each of the 32 temperature windows, each consisting of 10 000 MD steps with a time step of 0.002 ps). Figure 13A shows the heat capacity, which is broad with a small peak, similar to that of region II with this force field, indicating the absence of a sharp transition. Figure 13B shows the average rmsd of the helical segment (residues 277−289) as a function of temperature. At all temperatures, the average rmsd is about 6 Å from the crystal structure, indicating that helical conformations are not favored at any temperature with this force field. This observation is further confirmed by the helicity, which is less than 0.1 at all temperatures, indicating that no helical hydrogen bonds are present.

We calculated the free energy surface at 298 K (Figure 14) as a function of rmsd and helicity, revealing two distinct minima. The higher minimum is relatively narrow, and corresponds to structures about 2.8 Å from the crystal structure with helicities of 0.4−0.6, in good agreement with the values obtained in the MD simulations for this potential. This minimum corresponds to a state where the secondary structure observed in the crystal is partially retained. However, the global free energy minimum corresponds to structures that lie 5.5−7.5 Å from the crystal structure, with helicities below 0.1. This minimum is 4 kcal/mol lower in energy than the one associated with the crystal structure, and corresponds to extended and unfolded structures, in agreement with the unfolding observed in MD simulations for this potential. Our MD trajectories were not long enough to escape completely from the higher minimum.

We repeated the REMD simulations with CHARMM27/GBMV starting from the crystal structure. The simulations were performed for 5.0 μs (7805 replica cycles for each of 32 temperature windows, 10 000 MD steps for each with a time step of 0.002 ps). Calculated observables are shown in Figure 15. Compared to the CHARMM19/EEF1 force field for this region (Figure 13), and both force fields for region II, the heat capacity curve is sharp (Figure 15A), indicating a more well-defined transition. There is a clear signature for this transition in both the rmsd (Figure 15B) and helicity (Figure 15C), indicating melting as the temperature increases. The rmsd is higher (by 1 Å) and the helicity is lower (by about two hydrogen bonds) at 350 K for the REMD simulations with CHARMM27/GBMV compared to the MD runs with the same force field, indicating that the MD results are consistent, but did not reach equilibrium.

The free energy surface projected onto rmsd and helicity obtained with CHARMM27/GBMV has two comparable free energy minima, separated by a small barrier of 1 kcal/mol (Figure 16). One minimum spans conformations that are close to the crystal structure (rmsd about 1.5−3.3 Å), with helicity between 0.6 and 0.8. The other minimum consists of partially unfolded conformations that are between 4.5 and 5.5 Å from the crystal structure and have a helicity of around 0.4. This value still corresponds to four helical hydrogen bonds, which is significant for a small helix, indicating that there is still appreciable helical character. Inspection of

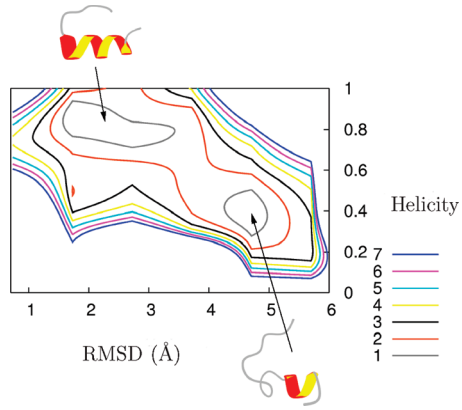


Figure 16. Free energy surface (kcal/mol) of region V with CHARMM27/GBMV as a function of rmsd and helicity at 298 K.

conformations from this minimum confirms that they have several helical turns. Hence the CHARMM27/GBMV force field supports an α-helical structure for both the region II and V peptides considered in the present work.

In summary, both CHARMM/EEF1 and CHARMM/GBMV support free energy minima corresponding to helical conformations close to the crystal structure for region V. However, for CHARMM/EEF1 the corresponding state is predicted to be unstable relative to extended structures, while for CHARMM/GBMV an equilibrium with partially unfolded helical conformations would be expected. The GROMOS96/SPC simulations did not produce any helical structure.

4. Conclusions

We have studied two evolutionarily conserved peptide fragments from the p53 DNA binding domain, both containing well-defined secondary structure in the intact protein. We used a number of force fields and solvent models, which produce somewhat different predictions for the structure propensity and stability. The region II peptide folds to its near-native β hairpin with the GROMOS96 force field and SPC explicit solvent. It is also stable in its hairpin conformation in MD runs with CHARMM19/EEF1 at 300 K, but forms a helical turn at the N-terminus for CHARMM27/GBMV at 350 K, which persists at higher temperatures where the hairpin is lost.

REMD simulations for region II with CHARMM19/EEF1 produce a free energy minimum containing a wide range of β conformations that are different from the crystal structure, perhaps suggesting a propensity for aggregation. With CHARMM27/GBMV the most stable conformations are helical in equilibrium for physiological temperatures, indicating that the MD trajectories were trapped in higher energy configurations. All the low-lying conformations for this potential have a well-defined helix in the middle of the sequence, with a varying degree of helicity at the N-terminus. The predictions of helical structure from CHARMM27/GBMV and homology modeling, which contrast with the β hairpin seen in the crystal structure and with the other potentials, suggest that the secondary structure preference of region II in the complete protein may be significantly affected by the environment.

For region V defined by residues 272−289 the consensus of homology models is that this peptide would lack secondary structure, although two of the models classify it as helical. In molecular dynamics simulations with CHARMM19/EEF1 and higher temperature runs for CHARMM27/GBMV the helix loses some of its helical structure. Simulations for region V with GROMOS96/SPC produced a non-native hairpin.

REMD simulations of region V for CHARMM19/EEF1 indicate that a variety of β-like conformations are present at physiological temperatures, as for region II with this potential. In contrast, for CHARMM27/GBMV the free energy minima below the folding transition temperature correspond to a range of helical structures that exhibit various degrees of unfolding compared to the crystal structure. Although there is a high energy minimum corresponding to the crystal structure for CHARMM19/EEF1, only the CHARMM27/GBMV combination predicts that the secondary structure observed in the crystal structure might be partially retained in the isolated peptide. The MD simulation in explicit water for GROMOS96/SPC resulted in collapse to a hairpin structure. These results again suggest that the environment in the complete protein may help to stabilize the secondary structure observed in the crystal. However, a longer sequence including additional residues associated with the extended helix observed in the NMR structure[21] would probably exhibit more helical character in isolation.

We should not necessarily expect small peptide fragments of larger proteins to exhibit stable secondary structure, and the present study reaffirms previous work[56,86] that reports a lack of consensus among different bimolecular force fields for such systems. It is well known that all force fields have some inherent bias. For example, CHARMM19/EEF1 may be less reliable at higher temperatures, because the experimental values used for specific heat capacities and enthalpies to parametrize this force field correspond to 300 K.[67] In this force field the partial charges are adjusted[67] so that the total charges of all residues are zero, and it is unclear what effect this will have on the electrostatics of sequences such as region V, which involve significant charges.

The generalized Born model has been reported to overstabilize secondary structure,[87] which may contribute to the formation of two distinct minima on the free energy surfaces of regions II and V separated by small barriers. It has also been reported that the GROMOS96 force field favors β hairpins,[56] which could explain why region II folds to a β hairpin comparable to the crystal structure, while homology models classify it as helical, and why region V also collapses to a β hairpin. We therefore conclude that the simulation results with different force fields should be considered as suggestions of how these systems might behave, which could be useful in interpreting future experiments. Finally, we note that intrinsically disordered regions[26] of proteins such as p53 may provide an even more stringent test of different force fields in computer simulations.

References

(1) Lane, D. P.; Crawford, L. V. *Nature (London)* **1979**, *278*, 261–263.

(2) Oren, M.; Rotter, V. *Cell. Mol. Life Sci.* **1999**, *55*, 9–11.

(3) May, P.; May, E. *Oncogene* **1999**, *18*, 7621–7636.

(4) Vogelstein, B.; Lane, D.; Levine, A. J. *Nature (London)* **2000**, *408*, 307–310.

(5) *25 Years of p53 Research*; Hainaut, P., Wiman, K. G., Eds.; Springer: New York, 2005.

(6) Linzer, D. I.; Levine, A. J. *Cell* **1979**, *17*, 43–52.

(7) Soussi, T.; Dehouche, K.; Beroud, C. *Hum. Mutat.* **2000**, *15*, 105–113.

(8) Römer, L.; Klein, C.; Dehner, A.; Kessler, H.; Buchner, J. *Angew. Chem., Int. Ed.* **2006**, *45*, 6440–6460.

(9) Joeger, A. C.; Fersht, A. R. *Oncogene* **2007**, *26*, 2226–2242.

(10) Freedman, D. A.; Wu, L.; Levine, A. J. *Cell. Mol. Life Sci.* **1999**, *55*, 96–107.

(11) Sherr, C. J. *Genes Dev.* **1998**, *12*, 2984–2991.

(12) Dasika, G. K.; Lin, S. C.; Zhao, S.; Sung, P.; Tomkinson, A.; Lee, E. Y. *Oncogene* **1999**, *18*, 7883–7899.

(13) Bartek, J.; Lukas, J. *Curr. Opin. Cell. Biol.* **2001**, *13*, 738–747.

(14) Salomoni, P.; Pandolfi, P. P. *Cell* **2002**, *108*, 165–170.

(15) Colombo, E.; Marine, J. C.; Danovi, D.; Falini, B.; Pelicci, P. G. *Nat. Cell Biol.* **2002**, *4*, 529–533.

(16) Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Lecine, A. J.; Pavletich, N. P. *Science* **1996**, *274*, 948–953.

(17) Hupp, T. R.; Lane, D. P.; Ball, K. L. *Biochem. J.* **2000**, *352*, 1–17.

(18) Cho, Y.; Gorina, S.; Jeffrey, P. D.; Pavletich, N. P. *Science* **1994**, *265*, 346–355.

(19) Weinberg, R. L.; Veprintsev, D. B.; Fersht, A. R. *J. Mol. Biol.* **2004**, *341*, 1145–1159.

(20) Tidow, H.; Melero, R.; Mylonas, E.; Freund, S. M. V.; Grossmann, J. G.; Carazo, J. M.; Svergun, D. I.; Valle, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 12324–12329.

(21) Nadillas, J. M. P. C.; Tidow, H.; Freund, S. M. V.; Rutherford, T. J.; Ang, H. C.; Fersht, A. F. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2109–2114.

(22) Rustandi, R. R.; Baldisseri, D. M.; Weber, D. J. *Nat. Struct. Biol.* **2000**, *7*, 570–574.

(23) Clore, G. M.; Ernst, J.; Clubb, R.; Omichinski, J. G.; Kennedy, W. M. P.; Sakaguchi, K. *Nat. Struct. Biol.* **1995**, *2*, 321–333.

(24) Lee, W.; Harvey, T. S.; Yin, Y.; Yau, P.; Litchfield, D.; Arrowsmith, C. H. *Nat. Struct. Biol.* **1994**, *1*, 877–890.

(25) Dawson, R.; Muller, L.; Dehner, A.; Klein, C.; Kessler, H.; Buchner, J. *J. Mol. Biol.* **2003**, *332*, 1131–1141.

(26) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762–5767.

(27) el Deiry, W. S.; Kern, S. E.; Pietenpol, J. A.; Kinzler, K. W.; Vogelstein, B. *Nat. Genet.* **1992**, *1*, 45–49.

(28) Kern, S. E.; Kinzler, K. W.; Bruskin, A.; Jarosz, D.; Friedman, P.; Prives, C.; Vogelstein, B. *Science* **1991**, *252*, 1708–1711.

(29) Kitayner, M.; Rozenberg, H.; Kessler, N.; Rabinovich, D.; Shaulov, L.; Haran, T. E.; Shakked, Z. *Mol. Cell* **2006**, *22*, 741–753.

(30) Ho, W. C.; Fitzgerald, M. X.; Marmorstein, R. *J. Biol. Chem.* **2006**, *281*, 20494–20502.

(31) Sionov, R. V.; Haupt, Y. *Oncogene* **1999**, *18*, 6145–6157.

(32) Reihsaus, E.; Kohler, M.; Kraiss, S.; Oren, M.; Montenarh, M. *Oncogene* **1990**, *5*, 137–145.

(33) Barak, Y.; Juven, T.; Haffner, R.; Oren, M. *EMBO J.* **1993**, *12*, 461–468.

(34) Momand, J.; Zambetti, G. P.; Olson, D. C.; George, D.; Levine, A. J. *Cell* **1992**, *69*, 1237–1245.

(35) Uesugi, M.; Verdine, G. L. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14801–14806.

(36) Yu, G. W.; Rudiger, S.; Veprintsev, D.; Freund, S.; Fernandez-Fernandez, M. R.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1227–1232.

(37) Honda, R.; Tanaka, H.; Yasuda, H. *FEBS Lett.* **1997**, *420*, 25–27.

(38) Li, M.; Brooks, C. L.; Wu-Baer, F.; Chen, D.; Baer, R.; Gu, W. *Science* **2003**, *302*, 1972–1975.

(39) Bullock, A. N.; Fersht, A. R. *Nat. Cancer Rev.* **2001**, *1*, 68–76.

(40) Roth, J. A.; Nguyen, D.; Lawrence, D. D.; Kemp, B. L.; Carrasco, C. H.; Ferson, D. Z.; Hong, W. K.; Komaki, R.; Lee, J. J.; Nesbitt, J. C.; Pisters, K. M. W.; Putnam, J. B.; Schea, R.; Shin, D. M.; Walsh, G. L.; Dolormente, M. M.; Han, C. I.; Martin, F. D.; Yen, N.; Xu, K.; Stephens, L. C.; McDonnell, T. J.; Mukhopadhyay, T.; Cai, D. *Nat. Med.* **1996**, *2*, 985–991.

(41) Seth, P.; Katayose, D.; Li, Z. W.; Kim, M.; Wersto, R.; Shanmugam, N.; Ohri, E.; Mudahar, B.; Rakkar, A. N. S.; Kadoli, P.; Cowan, K. *Cancer Gene Ther.* **1997**, *4*, 383–390.

(42) Kawabe, S.; Munshi, A.; Zumstein, L. A.; Wilson, D. R.; Roth, J. A.; Meyn, R. E. *Int. J. Radiat. Biol.* **2001**, *77*, 185–194.

(43) Kataoka, M.; Schumacher, G.; Cristiano, R. J.; Atkinson, E. N.; Roth, J. A.; Mukhopadhyay, T. *Cancer Res.* **1998**, *58*, 4761–4765.

(44) Nielsen, L. L.; Dell, J.; Maxwell, E.; Armstrong, L.; Maneval, D.; Catino, J. J. *Cancer Gene Ther.* **1997**, *4*, 129–138.

(45) Spitz, F. R.; Nguyen, D.; Skibber, J. M.; Cusack, J.; Roth, J. A.; Cristiano, R. J. *Anticancer Res.* **1996**, *16*, 3415–3422.

(46) Kubbutat, M. H. G.; Jones, S. N.; Vousden, K. H. *Nature (London)* **1997**, *387*, 299–303.

(47) Haupt, Y.; Maya, R.; Kazaz, A.; Oren, M. *Nature (London)* **1997**, *387*, 296–299.

(48) Wasylyk, C.; Salvi, R.; Argentini, M.; Dureuil, C.; Delumeau, I.; Abecassis, J.; Debussche, L.; Wasylyk, B. *Oncogene* **1999**, *18*, 1921–1934.

(49) Pavletich, N. P.; Chambers, K. A.; Pabo, C. O. *Genes Dev.* **1993**, *7*, 2556–2564.

(50) Bullock, A. N.; Henckel, J.; Fersht, A. R. *Oncogene* **2000**, *19*, 2000.

(51) Nikolova, P. V.; Henckel, J.; Lane, D. P.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14675–14680.

(52) Joerger, A. C.; Allen, M. D.; Fersht, A. R. *J. Biol. Chem.* **2004**, *279*, 1291–1296.

(53) Ang, H. C.; Joerger, A. C.; Mayer, S.; Fersht, A. R. *J. Biol. Chem.* **2006**, *281*, 21934–21941.

(54) Eisenmenger, F.; Hansmann, U. H. E. *J. Phys. Chem. B* **1997**, *101*, 3304–3310.

(55) Lwin, T. Z.; Luo, R. *Protein Sci.* **2006**, *15*, 2642.

(56) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2004**, *386*, 460–467.

(57) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(58) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(59) Oostenbrink, C.; Villa, A.; Mark, A. E.; Gunsteren, W. F. V. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(60) Berendsen, H. J.; Postma, J. P.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; D. Reidel Publishing Co.: Dordrecht, 1981; pp 331−342.

(61) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(62) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(63) Darden, T.; York, D.; Pedersn, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(64) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577–8592.

(65) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(66) Brooks, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. Charmm: The energy function and its parameterization with an overview of the program. In *The Encyclopedia of Computational Chemistry*; John Wiley and Sons: Chichester, 1998; pp 271−277.

(67) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.

(68) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(69) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.

(70) Inuzuka, Y.; Lazaridis, T. *Proteins* **2000**, *41*, 21–32.

(71) Hassan, S. A.; Mehler, E. L. *Proteins* **2002**, *47*, 45–61.

(72) Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1617–1632.

(73) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990–2001.

(74) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(75) MacKerell, A. D.; Feig, M.; Brooks, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(76) Karanicolas, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3432–3437.

(77) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(78) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(79) Geyer, C. J. *Proceedings of the 23rd symposium on the interface*; Computing Science and Statistics Interface Foundation of North America: Fairfax Station, VA, 1991.

(80) Hansmann, U. H. E.; Okamoto, Y. *Physica A* **1994**, *212*, 415–437.

(81) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.

(82) Feig, M.; Karanicolas, J.; Brooks, C. L. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.

(83) Sasai, M.; Ohmine, I.; Ramaswamy, R. *J. Chem. Phys.* **1991**, *96*, 3045–3053.

(84) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(85) Wallner, B.; Elofsson, A. *Protein Sci.* **2005**, *14*, 1315–1327.

(86) Bursulaya, B.; Brooks, C. L. *J. Phys. Chem. B.* **2002**, *104*, 12378–12383.

(87) Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J. *Bioinformatics* **1998**, *14*, 892–893.

(88) McGuffin, L. J.; Bryson, K.; Jones, D. T. *Bioinformatics* **2000**, *16*, 404–405.

(89) Garnier, J.; Gibrat, J. F.; Robson, B. *Methods Enzymol.* **1996**, *266*, 540–553.

(90) Guermeur, Y. Ph.D. Thesis, Université Paris 6, Paris, France, 1997.

(91) Kneller, D. G.; Cohen, F. E.; Langridge, R. *J. Mol. Biol.* **1990**, *214*, 171–182.

(92) Pollastri, G.; McLysaght, A. *Bioinformatics* **2004**, *21*, 1719–1720.

(93) Geourjon, C.; Delage, G. *Comput. Appl. Biosci.* **1995**, *11*, 681–684.

(94) Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. *Bioinformatics* **2006**, *22*, 195–201.

(95) Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. *Proteins* **2002**, *47*, 228–235.

(96) Wang, P.; Matsukawa, S.; Kameda, T.; Kurosu, H.; Ando, I. *J. Mol. Struct.* **1997**, *435*, 229–234.

(97) Imamura, H.; Chen, J. Z. Y. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 459–468.

CT8005387

JCTC Journal of Chemical Theory and Computation

Thermostat Artifacts in Replica Exchange Molecular Dynamics Simulations

Edina Rosta,[†] Nicolae-Viorel Buchete,[‡] and Gerhard Hummer*,[†]

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, and School of Physics, University College Dublin, Belfield, Dublin 4, Ireland

Received December 15, 2008

Abstract: We explore the effects of thermostats in replica exchange molecular dynamics (REMD) simulations. For thermostats that do not produce a canonical ensemble, REMD simulations are found to distort the configuration-space distributions. For bulk water, we find small deviations of the average potential energies, the buildup of tails in the potential energy distributions, and artificial correlations between the energies at different temperatures. If a solute is present, as in protein folding simulations, its conformational equilibrium can be altered. In REMD simulations of a helix-forming peptide with a weak-coupling (Berendsen) thermostat, we find that the folded state is overpopulated by about 10% at low temperatures, and underpopulated at high temperatures. As a consequence, the enthalpy of folding deviates by almost 3 kcal/mol from the correct value. The reason for this population shift is that noncanonical ensembles with narrowed potential energy fluctuations artificially bias toward replica exchanges between low-energy folded structures at the high temperature and high-energy unfolded structures at the low temperature. We conclude that REMD simulations should only be performed in conjunction with thermostats that produce a canonical ensemble.

1. Introduction

Replica exchange molecular dynamics (REMD)[1,2] is a widely used method to enhance the conformational sampling of molecular dynamics (MD) simulations.[3] In typical REMD simulations, several "replicas" i (i.e., copies of a physical system) are simulated in parallel at different temperatures T_i. At regular intervals, attempts are made to exchange the structures of different replicas to increase the conformational sampling efficiency at the lower temperatures.[1] The exchange of a structure X of a replica at temperature T_1 with a structure Y of a replica at temperature T_2 is accepted with probability $p^{\mathrm{acc}}(XY \rightarrow YX) = \min\{1, \exp(\Delta\beta\Delta U)\}$, where $\Delta U = U(Y) - U(X)$ is the potential energy difference, and $\Delta\beta = \beta_2 - \beta_1$ with $\beta_i^{-1} = k_B T_i$ and k_B being Boltzmann's constant. This acceptance criterion is designed to maintain canonical probability distributions in configuration space, $p_i(X) \propto e^{-\beta_i U(X)}$, at each temperature T_i, as follows from the detailed balance relation

$$\frac{p^{\mathrm{acc}}(XY \rightarrow YX)}{p^{\mathrm{acc}}(YX \rightarrow XY)} = \frac{p_1(Y)p_2(X)}{p_1(X)p_2(Y)} = e^{\Delta\beta\Delta U} \tag{1}$$

After an accepted exchange, particle velocities can be reassigned from a Maxwell−Boltzmann distribution at the new temperatures T_1 and T_2, respectively. Alternatively, the old velocities can be scaled by factors $(T_1/T_2)^{1/2}$ and $(T_2/T_1)^{1/2}$, respectively.[1,2]

Here, we explore the effects of combining REMD with thermostats that do not produce canonical ensembles. The weak-coupling (W−C) thermostat[4] (often referred to as "Berendsen" thermostat) is widely used in biomolecular simulations because of its stability and efficiency, but produces a noncanonical phase-space distribution.[5−8] As a consequence, the detailed balance relation, eq 1, is not satisfied. Problems with ergodicity in W−C simulations combined with REMD were identified previously in an

* To whom correspondence should be addressed. E-mail: Gerhard.Hummer@nih.gov.

[†] National Institutes of Health.

[‡] University College Dublin.

10.1021/ct800557h CCC: $40.75
Published on Web 04/09/2009

insightful study by Cooke and Schmidler.[9] We show that for bulk water at near-ambient conditions the effects of using a W−C thermostat in REMD simulations are relatively small. In contrast, the folding/unfolding equilibrium of a small peptide in water is shifted significantly by the thermostat, resulting in an overpopulation of the folded state at low temperatures and an underpopulation at high temperatures. The reason for this population shift is that the narrowed potential energy distributions in the noncanonical ensembles artificially favor replica exchanges between low-energy folded structures of the high-temperature replica and high-energy unfolded structures of the low-temperature replica.

We discuss possible ways to address the artificial population shift in REMD simulations with W−C thermostats. In principle, the acceptance criterion could be adjusted to maintain the energy distributions created by the thermostat, as implemented previously for microcanonical dynamics.[10] In practice, prior knowledge about the energy distributions, including their tails, is normally not available. We thus conclude that REMD simulations with a standard acceptance criterion, eq 1, should only be performed in conjunction with thermostats that preserve a canonical ensemble, such as Langevin thermostats,[11] Andersen thermostats,[12] Nosé−Hoover chains[13,14] based on the Nosé−Hoover thermostat,[15,16] or hybrid Monte Carlo.[17]

2. Methods

2.1. Water Simulations. MD simulations are widely used to study water and aqueous solutions. Even for relatively large solutes such as proteins, the main contribution to the heat capacity of the system, and thus to the potential energy fluctuations, typically comes from the water solvent. To explore the effects of combining REMD with thermostats, we thus first study bulk TIP3P water[18] at near-ambient conditions with a particle density of $\rho = 33.0$ nm^{-3} and at temperatures of $T = 300$ and 310 K. We compare the results of regular equilibrium MD and REMD simulations using a Langevin thermostat[11] with a collision frequency of 2 ps^{-1}, W−C thermostats[4] with coupling constants of 0.5 and 5 ps, and simulations run at constant energy (NVE ensemble; only regular MD). In all simulations, the system contains 1024 water molecules in a cubic, periodically replicated box of constant volume. Both regular MD and REMD simulations are performed using the sander module of Amber 9[19] with a time step of 0.002 ps, particle-mesh Ewald (PME) summation[20] with a grid width <1 Å, a real-space cutoff of 10 Å, and production times of 5 ns for each combination of thermostat and temperature, including the REMD runs. Two replicas are used in the REMD runs at temperatures of 300 and 310 K, respectively, with replica exchange attempts every 1 ps. In the NVE runs, the average kinetic energies correspond to temperatures of 299.9 and 310.1 K, respectively.

2.2. Protein Folding. To quantify the thermostat effects on the folding/unfolding equilibrium of a protein, we performed simulations of a helix-forming peptide Ala_5 in water.[3,21] In these simulations, the GROMACS molecular dynamics package[22] was used to run both standard MD (GROMACS version 3.3.0) and REMD simulations (GROMACS version 3.3.1) of the folding of blocked Ala_5(CH_3CO-Ala_5-$NHCH_3$) in TIP3P water.[18] For the peptide, we used the AMBER-GSS force field[23] ported to GROMACS.[24] The simulations were performed with periodic boundary conditions and PME electrostatics[20] using a real-space cutoff distance of 10 Å and a grid width <1 Å. REMD simulations were performed both with a W−C thermostat (coupling constant of 1 ps)[4] and a Langevin thermostat (collision frequency of 1 ps^{-1}).[22] Standard MD simulations were performed with a W−C thermostat. The pressure was held constant at $p = 1$ bar using a W−C barostat with a coupling constant of 5 ps.[4] The replica-exchange acceptance criterion was adjusted for simulations in an NPT ensemble by replacing the energy difference ΔU in eq 1 with the enthalpy difference ΔH, where $H = U + pV$ with V the fluctuating system volume. Since the compressibility of water is low near ambient conditions and pressure has little effect on the helix−coil equilibrium,[25] we expect that any deviations between MD and REMD simulations are caused primarily by the use of a W−C thermostat in the REMD runs and to a lesser degree by using a barostat that produces non-Boltzmann enthalpy distributions. A time step of 2 fs was used in conjunction with constrained bonds of hydrogen atoms.[26] The simulation box contained 1050 TIP3P water molecules.[18] Four independent trajectories starting from different initial conformations were created for each of the three setups (MD/W−C, REMD/W−C, and REMD/Langevin, respectively). The standard MD runs[21] were performed for 4 × 250 ns at 300 and 350 K, and 4 × 200 ns at 310, 325, and 340 K. REMD simulations of 150 ns duration (per replica) were run for each thermostat and each of the four initial conditions. We used 12 replicas spanning the 295−350 K temperature range, for a combined simulation time of 4 × 150 ns at each temperature.[3] Coordinates were saved every 1 ps and REMD exchanges were attempted every 5 ps.

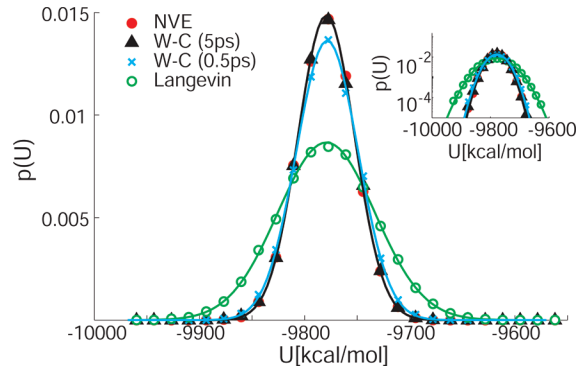


Figure 1. Potential energy distributions for simulations of TIP3P water at ambient conditions using Langevin (green circles) and W−C thermostats[4] with coupling constants of 0.5 ps (black triangles) and 5 ps (blue crosses), and in an NVE simulation (red dots). The continuous lines show Gaussians of corresponding means and variances. The inset shows the same distributions on a semilogarithmic scale.

3. Results and Discussion

3.1. Bulk Water. *MD.* Figure 1 shows the potential energy distributions for bulk TIP3P water obtained from regular MD with W−C and Langevin thermostats, and in

Table 1. Potential Energy Distributions Using Constant Energy (NVE) Dynamics, Langevin Dynamics, and W−C Thermostats for Bulk Water[a]

			300 K			310 K				
	thermostat	τ[ps]	$\langle U\rangle$	σ^2	skew	$\langle U\rangle$	σ^2	skew	g	$C_{12}[10^{-4}]$
	NVE		−9779.3(2)	709(7)	−0.01(2)	−9665.7(2)	757(8)	−0.01(2)	2.9	−2(3)
MD	Langevin	2.0	−9779.3(7)	2120(30)	0.00(2)	−9665.8(7)	2110(30)	−0.01(2)	1.0	−3(3)
	weak coupling	5.0	−9779.4(4)	722(8)	−0.03(2)	−9665.9(3)	767(7)	−0.01(2)	2.8	0(3)
	weak coupling	0.5	−9778.5(4)	858(9)	−0.01(2)	−9665.5(4)	900(10)	−0.01(2)	2.4	−2(3)
REMD	Langevin	2.0	−9777.8(7)	2110(30)	−0.01(2)	−9666.7(8)	2150(30)	−0.01(2)	1.0	−1(2)
	weak coupling	5.0	*−9773.4(7)*	*910(10)*	*0.26(3)*	*−9671.2(7)*	*940(10)*	*−0.31(2)*	*2.0*	*−55(3)*
	weak coupling	0.5	*−9775.2(5)*	*990(10)*	*0.17(2)*	*−9669.0(5)*	*1000(10)*	*−0.18(3)*	*2.0*	*−42(4)*

[a] Energies are in units of kcal/mol. The numbers in parentheses are the estimated statistical errors in the last digits (one standard deviation). Italic numbers indicate deviations between REMD and MD results that exceed twice the combined statistical errors. τ is the thermostat coupling time. The skew coefficient is defined as skew $= \langle(U-\langle U\rangle)^3\rangle/\langle(U-\langle U\rangle)^2\rangle^{3/2}$. $g = C_V/C_V^{\text{fluc}}$ is the ratio of the excess heat capacities calculated from the temperature derivative and the canonical fluctuation formula. C_{12} is the normalized cross-correlation coefficient.

the NVE ensemble. We find that at a given temperature all distributions are well approximated by Gaussians of nearly identical means. However, the widths of the distributions vary considerably, with the W−C and NVE simulations producing narrower distributions than the Langevin simulations, consistent with results from earlier studies.[5−7] As shown in Table 1, the variances in the potential energy distributions differ by up to a factor of 3.

To test whether the potential energy distributions are consistent with a canonical distribution, we compare the excess heat capacity at constant volume calculated (1) from the temperature derivative of the average potential energy, $C_V = \partial\langle U\rangle/\partial T$, and (2) from the fluctuations in the potential energy, $C_V^{\text{fluc}} = \sigma^2/(k_B T^2)$ where $\sigma^2 = \langle U^2\rangle - \langle U\rangle^2$ is the variance in the potential energy. For a canonical ensemble, the two expressions are identical, $C_V \equiv C_V^{\text{fluc}}$. However, we will show in the following that in the noncanonical ensembles created by W−C thermostats, C_V^{fluc} differs from C_V because of the narrowed potential energy distributions. To estimate the temperature derivative in the expression for C_V, we use the difference between the average potential energies at 310 and 300 K, $C_V \approx (\langle U\rangle_2 - \langle U\rangle_1)/(T_2 - T_1)$. The variance in C_V^{fluc} is averaged over the two temperatures, $\sigma^2 \approx (\sigma_1^2 + \sigma_2^2)/2$. As listed in Table 1, we find that the ratio of the two expressions for the excess heat capacity, $g = C_V/C_V^{\text{fluc}}$, is 1.0 for the Langevin simulations, consistent with canonical distributions. In contrast, for the W−C and NVE simulations g varies between 2.4 and 2.9, indicative of strong deviations from the canonical distribution.

REMD. Figure 2 compares the potential energy distributions of bulk water at 300 and 310 K obtained from MD and REMD simulations using W−C thermostats with a coupling constant of 5 ps. The results for different thermostats are summarized in Table 1. We find that in REMD with a Langevin thermostat, the distributions do not change compared to those from MD simulations. In contrast, REMD significantly changes the potential energy distributions when a W−C thermostat is used: the mean energies at the two temperatures move together, the variances increase, and the distributions become skewed with pronounced non-Gaussian tails. Table 1 also lists the normalized cross-correlation coefficient $C_{12} = \langle(U_1 - \langle U_1\rangle)(U_2 - \langle U_2\rangle)\rangle/(\sigma_1\sigma_2)$ between the instantaneous potential energies of the two replicas. We find

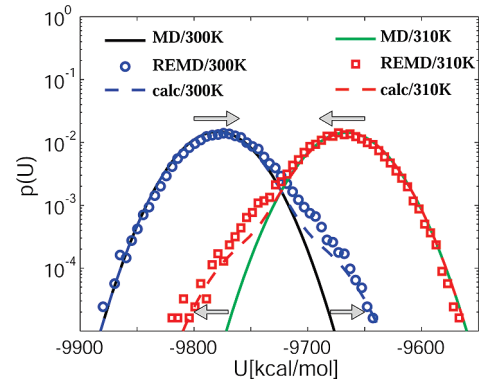


Figure 2. Comparison of bulk TIP3P water potential energy distributions from MD and REMD simulations (semilogarithmic scale). The blue circles and red squares show the REMD results for the replicas at 300 and 310 K, respectively. The blue and red dashed lines are calculated from iterated numerical solutions of eq 2 at 300 and 310 K, respectively. For reference, Gaussian approximations to the MD results using a W−C thermostat[4] with a coupling time of 5 ps are shown as black (300 K) and green lines (310 K). The arrows indicate the change in the energy distributions of REMD simulations compared to MD, in particular the buildup of artificial tails and the small shifts in the means.

that in the REMD simulations with a Langevin thermostat, the energies are uncorrelated ($C_{12} = 0$). In REMD with a W−C thermostat, we find small but significant correlations, $C_{12} \approx -0.0050 \pm 0.0004$. We conclude from these results that REMD with a W−C thermostat can significantly alter the potential energy distributions.

Model of Thermostat Effects in REMD. The effects of thermostats in conjunction with REMD can be understood from a simple model, similar to the ones used previously[27,28] to study the efficiency of REMD. With $p_i^{\text{MD}}(U)$ the potential energy distribution at temperature T_i, the joint probability distribution of the potential energies U and W at the two temperatures T_1 and T_2 is $p^{\text{MD}}(U,W) = p_1^{\text{MD}}(U)p_2^{\text{MD}}(W)$. Let us consider how replica exchange alters this distribution. After an attempted replica exchange, a given pair of energies U and W is obtained either from an accepted move starting with W and U or from a rejected move starting with U and W. Replica exchange thus transforms $p^{\text{MD}}(U,W)$ to:

$$p^{\mathrm{REMD}}(U,W) = p^{\mathrm{MD}}(W,U)p^{\mathrm{acc}}(W,U) + p^{\mathrm{MD}}(U,W)[1 - p^{\mathrm{acc}}(U,W)] \quad (2)$$

where $p^{\mathrm{acc}}(U,W)$ is the probability that a replica exchange is accepted if U and W are the potential energies of the structures in replicas 1 and 2, respectively. If the acceptance probabilities satisfy detailed balance, $p^{\mathrm{acc}}(U,W)p^{\mathrm{MD}}(U,W) = p^{\mathrm{acc}}(W,U)p^{\mathrm{MD}}(W,U)$, then replica exchange preserves the distribution, $p^{\mathrm{REMD}}(U,W) = p^{\mathrm{MD}}(U,W)$; otherwise, the distribution will be distorted. In our model, we iterate eq 2 to self-consistency and then calculate the marginal distributions by integration, $p_1^{\mathrm{REMD}}(U) = \int p^{\mathrm{REMD}}(U,W)\mathrm{d}W$ and $p_2^{\mathrm{REMD}}(W) = \int p^{\mathrm{REMD}}(U,W)\mathrm{d}U$. In the numerical calculations we approximate the potential energy distributions by Gaussians with means and variances taken from the MD results in Table 1. As shown in Figure 1, Gaussians provide excellent approximations to the actual distributions over the whole range of potential energies sampled in the MD simulations.

Figure 2 compares the potential energy distributions obtained by iteration of eq 2 to those of the REMD simulations. We find that the model can account almost quantitatively for the effects of the W−C thermostat on the energy distributions. The distribution at the low temperature is shifted upward and has a pronounced tail toward the higher energies. The energy distribution at the high temperature is modified in the opposite direction: it is shifted toward lower energies and has a tail skewed to the left. The cross-correlation coefficient C_{12} predicted by the model is about −0.09, close to the $C_{12} \approx -0.005$ obtained from the W−C REMD simulations (Table 1). The smaller correlations in the REMD simulations can be explained by the randomizing effect of the thermostatted MD runs between replica exchanges.

The distortions seen in Figure 2 arise because the potential energy distributions of the thermostat are inconsistent with eq 1. For the narrowed energy distributions produced by the W−C thermostat, as compared to the wider canonical distributions, replica exchange is relatively more likely to be accepted for conformations in the high-energy tail at the low temperature, and in the low-energy tail at the high temperature. We will show in the following that this bias can result in distortions of the conformational equilibrium of a solute.

3.2. Protein Folding. REMD is widely used to study the folding of peptides and proteins. Protein folding can often be described by using only two dominant states, with a folded state being enthalpically stabilized against an entropically favored unfolded state. In the following, we explore how simulations of protein folding are affected by replica exchange when W−C thermostats are used. We analyze the simulation data from refs 3 and 21 for blocked Ala_5 in water, which folds into a short helical peptide for the force field used.

Figure 3 compares the relative populations of the folded (helical) state obtained from long MD runs at 300 and 350 K, and from REMD simulations using Langevin and W−C thermostats, respectively, with 12 replicas equally spaced in temperature between 295 and 350 K. The folded state is defined as in refs 3 and 21. We find that the equilibrium populations of the folded state agree for standard MD simulations using a W−C thermostat and REMD simulations using a Langevin thermostat. In contrast, the folded populations obtained from REMD simulations using a W−C thermostat differ significantly from both the standard MD results and the REMD/Langevin results. In the REMD runs with the W−C thermostat, the population of the folded (helical) peptide is increased at the low temperatures and decreased at the high temperatures, with ∼10% shifts in the folded populations at 300 and at 350 K.

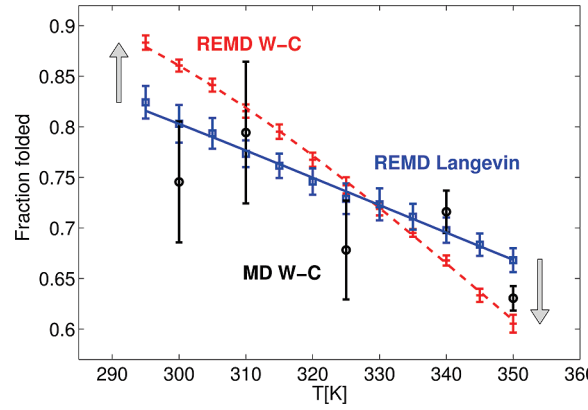


Figure 3. Relative populations of the folded (helical) state of Ala_5 as a function of temperature (red crosses: REMD with W−C thermostat; blue squares: REMD with Langevin thermostat; black circles: MD). The dashed line (REMD with W−C) and solid line (REMD with Langevin) are fitted melting profiles. The error bars correspond to one standard deviation estimated from four independent runs. In case of the MD simulations, the standard deviations agree well with the theoretical value obtained from the kinetic rate coefficients of folding and unfolding[21] in a two-state kinetic model.[29] Vertical arrows indicate the population changes in the REMD W−C simulations.

Table 2. Enthalpy, ΔH, and Entropy, ΔS, of Folding Obtained from Fits to the Melting Profiles (Figure 3)[a]

	ΔH [kcal/mol]	ΔS [cal/(mol K)]	$\langle U\rangle_f - \langle U\rangle_u$ [kcal/mol]
MD	−3.20 ± 1.00	−8.0 ± 2.9	−3.00 ± 0.10
REMD (Langevin)	−2.93 ± 0.26	−7.0 ± 0.8	−3.13 ± 0.10
REMD (W−C)	−5.74 ± 0.14	−15.5 ± 0.4	−3.68 ± 0.07

[a] The last column lists the difference in average total potential energy of the folded and unfolded states. The estimated statistical errors (one standard deviation) are also given.

For a quantitative comparison, we fit the relative populations $p_f(T)$ of the folded state to a melting profile $p_f(T) = 1/[1 + \exp(\Delta H/k_B T - \Delta S/k_B)]$ where ΔH and ΔS are the enthalpy and entropy of folding, respectively, which are assumed to be independent of temperature in the range of 295 to 350 K. The results are shown in Table 2. We find that both the enthalpy and entropy of folding from the Langevin-thermostatted REMD and the MD simulations are in excellent agreement. In contrast, both the enthalpy and entropy obtained from REMD/W−C and REMD/Langevin simulations differ by about 7 times their respective combined standard deviations. For the small peptide, the systematic error in the enthalpy of folding is almost 3 kcal/mol. We conclude from these inconsistencies that REMD with a W−C

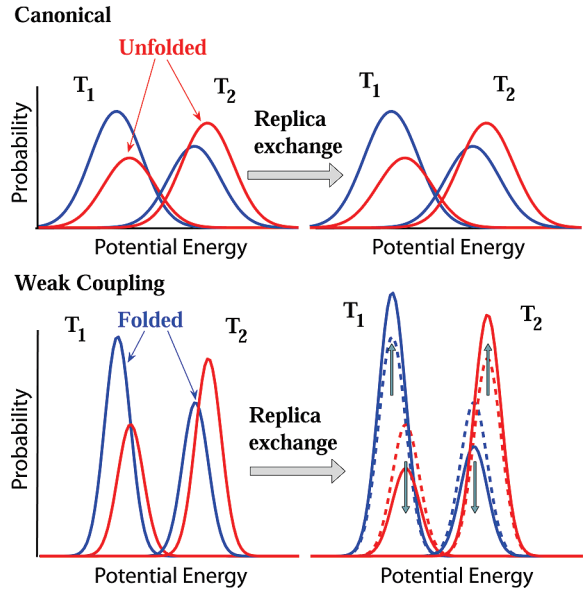


Figure 4. Schematic of the effect of REMD with W−C thermostats on folding probabilities. Potential energy probability distributions of folded (blue) and unfolded (red) states are shown at two temperatures before and after replica exchange. For classical canonical distributions (upper panel), REMD does not affect the distributions. However, for the narrowed distributions in W−C REMD (lower panel), replica exchange alters the relative populations of folded and unfolded states. Folded states become overpopulated at low temperatures and underpopulated at high temperatures, as indicated by the vertical arrows.

thermostat significantly alters the temperature dependence of the folded population.

Qualitatively, the change in the folded populations induced by combining REMD with a W−C thermostat can be explained by using a simple model, as illustrated in Figure 4. In this model, we assume that the folded state has lower enthalpy than the unfolded state. The blue (red) lines show the potential energy distribution of the folded (unfolded) states. In the canonical case, replica exchange leaves the distributions unchanged, thus preserving equilibrium. For a W−C thermostat (bottom), the distributions are narrower, reducing the overlap of the distributions at different temperatures. As a consequence, the probability of acceptance of replica exchange moves is reduced. In addition, in the W−C REMD there are relatively more accepted replica exchanges between the unfolded state at the lower temperature T_1 and the folded state at T_2. As a result, the population of the folded states artificially increases at T_1 and decreases at T_2. In Appendix A we implement such a model, and show that it can account for the observed population shifts. Interestingly, this model also explains the observation of Cooke and Schmidler[9] that in their simulations of alanine dipeptide in vacuum using REMD with a W−C thermostat, the $C7_{ax}$ configuration was not populated in the lowest-temperature replica: $C7_{ax}$ has higher energy than $C7_{eq}$ and, according to the model in Figure 4, should have an artificially low population in the replicas at low temperatures.

For further validation of our model, we also calculate ΔH directly from the difference in the average total potential energies of the system with the peptide in the folded and unfolded state, respectively (see Table 2; the small pV contribution is ignored). The potential energy differences between folded and unfolded states averaged over all temperatures and all four initial conditions are listed for both the MD and REMD simulations with Langevin and W−C thermostats. Errors are estimated from the $5 \times 4 = 20$ and $12 \times 4 = 48$ independent MD and REMD simulations, respectively. ΔH obtained from the potential energy differences in REMD/Langevin and in standard MD simulations are consistent with the values obtained from the fit of the melting profile. In contrast, for the REMD simulations with the W−C thermostat the energy difference and the value of ΔH are inconsistent, differing by ∼10 combined standard deviations. ΔH obtained from the potential energy distributions of the REMD W−C simulations is significantly closer to the values obtained from the MD and REMD Langevin data, as expected from our model illustrated in Figure 4.

4. Conclusions

We showed that thermostats can significantly affect the outcome of REMD simulations, consistent with the results of earlier studies.[9] W−C thermostats produce potential energy distributions that are narrower than those expected for a canonical ensemble. If a standard replica-exchange acceptance criterion is used, replica exchange will distort these distributions and, as a result, shift the configuration-space populations. For bulk TIP3P water[18] at near-ambient conditions, the effects are statistically significant but overall small. We found that in W−C simulations the mean potential energies are shifted, become artificially correlated, and develop distributions with enhanced tails. No such effects were seen in REMD simulations using a Langevin thermostat that preserves the canonical distribution.

For protein folding, REMD with noncanonical thermostats is expected to result in more pronounced effects. Simulations of a small helix-forming peptide using a W−C thermostat showed that the folding probabilities are artificially enhanced by ∼10% at the low temperatures and reduced at the high temperatures. For thermostats that produce a higher variance than the canonical energy distribution, the opposite effects are expected.

We have used simple models to estimate the effects of thermostats on the energy distributions and folding equilibria. As input, these models use the energy differences between folded and unfolded states, and the means and variances of the potential energy. We found that the thermostat-induced changes in the energy distributions and folding equilibria could be predicted nearly quantitatively, suggesting that the models can be used to assess and possibly correct W−C REMD simulation results.

For noncanonical simulations (e.g., at constant energy or with W−C thermostats), the replica-exchange acceptance criterion, eq 1, has to be modified.[10] Without exact explicit expressions for the configuration-space distribution, such modifications are not easily possible in practice. We therefore conclude that REMD simulations should be carried out with thermostats that preserve canonical distributions, such as Langevin thermostats,[11] Andersen thermostats,[12] Nosé−Hoover

chains[13,14] based on the Nosé−Hoover thermostat[15,16] (which by itself is not guaranteed to produce a canonical distribution),[9,30] or the hybrid Monte Carlo method.[17]

Acknowledgment. We thank Dr. Attila Szabo for many helpful and stimulating discussions and Dr. Nicolas Fawzi for his valuable comments. This research used the Biowulf Linux cluster at the National Institutes of Health (NIH), and the Irish Centre for High-End Computing (ICHEC). E.R. and G.H. were supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases, NIH.

Appendix: Protein Folding in REMD Simulations

We use a simple model to analyze the effects of using W−C thermostats in protein-folding REMD simulations. In this model, the protein is assumed to have two states, a low-energy folded state and a high-energy unfolded state, $U_f < U_u$, with relative populations $p_f(T_i)$ at temperature T_i. For simplicity, we first consider only two replicas at two temperatures, $T_1 < T_2$. There are four distinct folding states overall: *ff*, *fu*, *uf*, and *uu*, where the first position indicates the folding state at T_1 and the second at T_2. At equilibrium, the populations of the four states are p_{ff}, p_{fu}, p_{uf}, and p_{uu}, respectively. An accepted replica exchange switches the states $X \in \{f,u\}$ and $Y \in \{f,u\}$ at temperatures T_1 and T_2. After an exchange attempt, the probability distribution of the four states changes to p_{XY}^{REMD}

$$\begin{aligned} p_{ff}^{\mathrm{REMD}} &= p_{ff} \\ p_{fu}^{\mathrm{REMD}} &= p_{fu}[1 - p^{\mathrm{acc}}(fu \rightarrow uf)] + p_{uf}p^{\mathrm{acc}}(uf \rightarrow fu) \\ p_{uf}^{\mathrm{REMD}} &= p_{uf}[1 - p^{\mathrm{acc}}(uf \rightarrow fu)] + p_{fu}p^{\mathrm{acc}}(fu \rightarrow uf) \\ p_{uu}^{\mathrm{REMD}} &= p_{uu} \end{aligned} \quad (3)$$

where $p_{XY}^{\mathrm{REMD}} = p_{XY}$ in the canonical case. The acceptance probabilities p^{acc} of replica exchanges depend on the potential energy difference. For simplicity, we assume that the solvent-energy distributions at each temperature T_i are Gaussians, $g_i(U)$, with means $\langle U \rangle_i$ and standard deviations σ_i, independent of the state of the protein. We thus ignore the non-Gaussian tails of the distributions. The distributions of the potential energies at temperature T_i in the folded and unfolded states are then $p_i(U|f) = g_i(U - U_f)$ and $p_i(U|u) = g_i(U - U_u)$, respectively. If the solvent-energy fluctuations are fast relative to folding and unfolding, they can be integrated out. The probability of accepting an exchange $XY \rightarrow YX$ then becomes

$$p^{\mathrm{acc}}(XY \rightarrow YX) = \int \mathrm{d}U \int \mathrm{d}W p_1(U|X) p_2(W|Y) \min\{1, \mathrm{e}^{(\beta_2 - \beta_1)(W-U)}\} \quad (4)$$

This model can easily be extended to more than two replicas. As in the model for bulk water, the coupled equations corresponding to eq 3 can be iterated to self-consistency.

We applied this model to the REMD simulations of Ala_5 folding in water. To calculate the replica-exchange acceptance probabilities, we assumed a constant enthalpy difference of folding, ΔH, and linear temperature dependences of σ and $\langle U \rangle$. The parameters were extracted from MD simulations of Ala_5 using a W−C thermostat with a 1 ps coupling constant at 300 and 350 K, respectively. ΔH and the folding probabilities at different temperatures were determined from the folding and unfolding rates extracted from the MD simulations[3,21] using a two-state model and assuming an Arrhenius temperature dependence. As in the REMD simulations, we used 12 temperatures spaced equally between 295 and 350 K. The iterative scheme corresponding to eq 3 was applied by alternating exchange attempts between replicas at neighboring temperatures. Consistent with the REMD simulations, the converged model led to an increase in the folded population at low temperatures and a decrease at high temperatures. However, the magnitude of the effect was smaller, resulting in ∼5% changes in the folded population at the two extreme temperatures.

References

(1) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(2) García, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782–2787.

(3) Buchete, N. V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.

(4) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(5) D'Alessandro, M.; Tenenbaum, A.; Amadei, A. *J. Phys. Chem. B* **2002**, *106*, 5050–5057.

(6) Morishita, T. *J. Chem. Phys.* **2000**, *113*, 2976–2982.

(7) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U.K., 1987.

(8) Frenkel, D.; Smit, B. *Understanding Molecular Simulation. From Algorithms to Applications*; Academic Press: San Diego, CA, 2002.

(9) Cooke, B.; Schmidler, S. C. *J. Chem. Phys.* **2008**, *129*, 164112.

(10) Calvo, F.; Neirotti, J. P.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **2000**, *112*, 10350–10357.

(11) Pastor, R. W.; Brooks, B. R.; Szabo, A. *Mol. Phys.* **1988**, *65*, 1409–1419.

(12) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

(13) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(14) Tobias, D. J.; Martyna, G. J.; Klein, M. L. *J. Phys. Chem.* **1993**, *97*, 12959–12966.

(15) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511.

(16) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(17) Duane, S.; Kennedy, A. D.; Pendleton, B. J.; Roweth, D. *Phys. Lett. B* **1987**, *195*, 216–222.

(18) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(19) Case, D. A.; Cheatham, T., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(20) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(21) Buchete, N. V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

(22) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(23) Nymeyer, H.; García, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934–13939.

(24) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472–2493.

(25) Paschek, D.; Gnanakaran, S.; García, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6765–6770.

(26) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(27) Sindhikara, D.; Meng, Y. L.; Roitberg, A. E. *J. Chem. Phys.* **2008**, *128*, 024103.

(28) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15340–15345.

(29) Berezhkovskii, A. M.; Szabo, A.; Weiss, G. H. *J. Chem. Phys.* **1999**, *110*, 9145–9150.

(30) Tuckerman, M. E.; Liu, Y.; Ciccotti, G.; Martyna, G. J. *J. Chem. Phys.* **2001**, *115*, 1678–1702.

CT800557H

JCTC Journal of Chemical Theory and Computation

Evidence for Stabilization of DNA/RNA−Protein Complexes Arising from Nucleobase−Amino Acid Stacking and T-Shaped Interactions

Lesley R. Rutledge, Holly F. Durst, and Stacey D. Wetmore*

Department of Chemistry and Biochemistry, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta, Canada T1K 3M4

Received December 19, 2008

Abstract: The stacking and T-shaped interactions between the natural DNA or RNA nucleobases (adenine, cytosine, guanine, thymine, uracil) and all aromatic amino acids (histidine, phenylalanine, tyrosine, tryptophan) were investigated using ab initio quantum mechanical calculations. We characterized the potential energy surface of nucleobase−amino acid dimers using the MP2/6-31G*(0.25) method. The stabilization energies in dimers with the strongest interactions were further examined at the CCSD(T)/CBS level of theory. Results at the highest level of theory possible for these systems indicate that both stacking and T-shaped interactions are very close in magnitude to biologically relevant hydrogen bonds. Additionally, T-shaped interactions are as strong, if not stronger, than the corresponding stacking interactions. Our systematic consideration of the interaction energies in 485 possible combinations of monomers shows that a variety of these contacts are essential when considering the role of aromatic amino acids in the binding of proteins to DNA or RNA. This work also illustrates how our calculated binding strengths can be used by biochemists to estimate the magnitude of these noncovalent interactions in a variety of DNA/RNA−protein active sites.

Introduction

Interactions between DNA/RNA and protein building blocks are essential for fundamental cellular processes such as transcription, DNA replication, DNA damage repair, apoptosis, and gene expression.[1,2] Thus, understanding how proteins recognize specific DNA sequences or certain nucleobases is crucial for understanding the role of these contacts in a variety of biological processes and medical applications. For example, gene expression is regulated by protein switches that bind to specific DNA sequences,[3,4] which has led to proposals that DNA−protein interactions can be used to target numerous genetic diseases through rational drug design.[3,5]

In attempts to understand how proteins recognize nucleobase sequences, structural analyses of a large number of DNA/RNA−protein complexes have been performed.[6−8] These studies reveal that it is not possible to establish a simple set of rules for predetermining interactions between DNA and protein building blocks.[9,10] Additionally, proteins often readily undergo structural changes to accommodate different nucleobases,[10−12] where each substrate exploits unique active site interactions to promote binding.[10] Because of the complexity, noncovalent interactions have been proposed to govern these biologically crucial contacts.[3,13]

There has been debate in the literature regarding the relevance of aromatic amino acids (Figure 1) in base recognition,[13,14] which may be attributed to a poor understanding of the nature of aromatic−aromatic interactions. Additionally, identifying important aromatic−aromatic contacts in experimental crystal structures can be more challenging than other intermolecular interactions, such as (conventional) hydrogen bonding, which has well-defined structural characteristics.[13] However, a recent study searched 141 DNA−protein and 61 RNA−protein complexes found in the Protein Data Bank (PDB) and revealed that aromatic amino acids appear in nucleic acid binding sites more frequently than expected.[15] Indeed, 532 and 242 aromatic−

* Corresponding author. Fax: (403) 329-2057. E-mail: stacey.wetmore@uleth.ca.

10.1021/ct800567q CCC: $40.75
Published on Web 03/30/2009

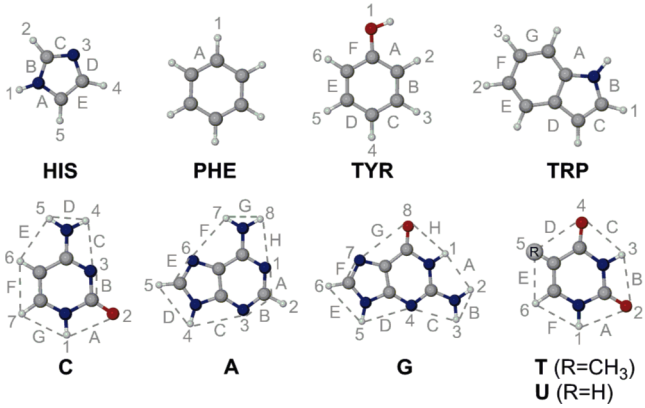


Figure 1. Structures and definition of edges of the aromatic amino acids (histidine (HIS), phenylalanine (PHE), tyrosine (TYR), tryptophan (TRP)) and the natural nucleobases (cytosine (C), adenine (A), guanine (G), thymine (T), uracil (U)).

Table 1. The Number of Aromatic−Aromatic Contacts between Amino Acids and DNA or RNA Bases Identified in the Protein Data Bank (PDB) by Baker and Grant[15]

	HIS		PHE		TYR		TRP	
	DNA	RNA	DNA	RNA	DNA	RNA	DNA	RNA
A	41	12	48	14	28	15	8	7
G	65	35	28	9	30	18	6	14
C	41	22	15	16	36	24	11	19
T	66	-	60	-	41	-	8	-
U	-	12	-	6	-	16	-	3

aromatic nucleobase−amino acid contacts were identified in DNA and RNA containing structures, respectively (Table 1).[15] Since these complexes involve all four aromatic amino acids, it can be proposed that all aromatic side chains play a vital role in binding nucleic acid bases.

Although identifying a large number of nucleic acid−protein aromatic−aromatic interactions is the first step for demonstrating their importance in biological processes, very little is understood about these contacts. For example, key unanswered questions include the following: how does the relative nucleobase−side chain orientation affect the strength, what is the nature behind these attractive forces, how strong are the contacts, how can these interactions be used in nature, do the aromatic amino acids play a role in protein complex stability and/or DNA/RNA recognition, and is there a simple way to predict the strength of these interactions in DNA/RNA−protein complexes observed in experimental crystal structures?

Computational modeling of aromatic−aromatic interactions can provide valuable insights into the above questions. Unfortunately, few accurate ab initio calculations have been conducted on stacked[16−20] or T-shaped[16,21] orientations between nucleobases and aromatic side chains. Adenine has been investigated by several groups since it is a fundamental building block commonly used to study protein structure.[22] For example, the Rooman group characterized T-shaped and stacked orientations between adenine and phenylalanine or histidine (neutral and protonated) using crystal structure geometries, as well as identified optimal geometries using MP2/6-31G*(0.8,0.2) calculations.[16] The Hu group also

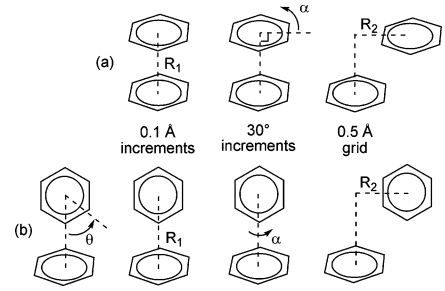


Figure 2. The definition of the geometric variables considered and increments used in (a) stacking potential energy surface scans and (b) T-shaped potential energy surface scans (angle of 'edge' rotation (θ), vertical separation (R_1), angle of rotation (α), and horizontal displacement (R_2)). See Figure 1 for θ edges considered.

carried out large scale data mining of the PDB to identify crystal structures involving adenine stacked with the aromatic amino acids (phenylalanine, tyrosine, and tryptophan), where their MP2/6-311+G(d) calculations provide evidence for strongly attractive forces between these $\pi-\pi$ systems.[19] Recently, the Tschumper group investigated adenine−phenylalanine stacked dimers in experimental crystal structure orientations, as well as fully optimized geometries, where CCSD(T)/CBS interaction energies were reported.[20] Despite their potential importance, to our best knowledge, a full study of the heterodimers of the aromatic amino acids and nucleobases in both stacked and T-shaped geometries has not yet been performed.

In this work, we present the first systematic investigation of both stacked and T-shaped interactions between all aromatic amino acids (histidine, phenylalanine, tyrosine, tryptophan) and natural nucleobases (adenine, cytosine, guanine, thymine, uracil) at the highest levels of accuracy possible for these systems. Specifically, we use MP2/6-31G*(0.25) to scan the potential energy surface of the nucleobase−amino acid dimers. Subsequently, the dimer orientations that yield the strongest interactions were further examined at the CCSD(T)/CBS level. Our unique methodology for scanning the potential energy surface reveals more information than simply identifying the important minima on these dimer surfaces. For example, this study also reveals how geometrical changes in the relative orientations of stacked and T-shaped dimers dictate the strength and importance of aromatic−aromatic contacts. This methodology also reveals significant information about the magnitude of these biologically relevant contacts by providing the most accurate comparison in the literature to date of the stacking and T-shaped interactions between all aromatic amino acids and DNA and RNA nucleobases.

Computational Methods

A series of MP2/6-31G*(0.25) single-point calculations were used to scan the potential energy surface for the stacked and T-shaped orientations of nucleobase−amino acid dimers according to several geometric variables (Figure 2), which is a similar approach to that used by Hobza and Šponer to study natural nucleobase dimers,[23−26] as well as our group to study dimers of the aromatic amino acids and natural or

damaged nucleobases.[18,21,27] Initially, C_s symmetric, MP2/6-31G(d) optimized monomers were placed so that the nucleobase and amino acid molecular planes were parallel (stacked) or perpendicular (T-shaped) to one another. In stacked dimers involving HIS, TYR, and TRP, two relative orientations of the molecular planes were considered. The first was obtained by stacking the amino acid and nucleobase orientations as shown in Figure 1, while the second conformation was obtained by mirror flipping the amino acid relative to Figure 1 before stacking with the nucleobase (indicated using a prime (′) throughout the manuscript). These orientations yield a total of 35 stacked complexes. T-Shaped dimers involving either amino acid edges or nucleobase edges were considered. To define the ring edge (monomer edge) directed toward the π-system (monomer face), the geometric variable θ (Figure 2b) is introduced. Our edge nomenclature (Figure 1) uses a number to indicate an atom directed toward the monomer face, and a letter to indicate a bridged structure involving more than one atom directed toward the π-system. For example, in dimers involving PHE edges, $\theta = 1$ indicates that a hydrogen atom from the PHE ring is directed at the π-system, while θ = A indicates that a C−C bond of the PHE ring is parallel to the nucleobase molecular plane. This choice of edges led to 170 amino acid edge and 280 nucleobase edge dimers. A full description of initial structures is provided in the computational methods in the Supporting Information.

From these initial structures, three geometric variables were considered in our potential energy surface scans (Figure 2): the vertical separation (R_1), the angle of rotation (α), and the horizontal displacement (R_2) between monomers. The increments used in our scans are indicated in Figure 2, and full details of our initial structures, as well as the (R_2) shift directions, are provided in the Supporting Information. We note that our R_2 calculations shift one monomer relative to the other monomer in two directions that are perpendicular to the R_1 axis. As a consequence of the large number of calculations involved in R_2 scans, we only considered the strongest 124 (out of 450) T-shaped dimers, while all 35 stacked dimers were further investigated.

The MP2 binding strengths for the orientations with the strongest (most negative) interactions as identified in potential energy surface scans were extrapolated to the CCSD(T) level at the complete basis set (CBS) limit. We considered all stacked dimers, the strongest amino acid edge dimers, the strongest nucleobase edge dimers that involve the model N−H glycosidic bond (relevant for nucleobases), and the strongest nucleobase edge dimers that do not involve the model glycosidic bond (relevant for nucleotides and nucleosides). MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ energies were used to extrapolate to the MP2 complete basis set limit (CBS) using the procedure of Helgaker.[28,29] To account for errors in MP2 electron correlation, the difference between the MP2/6-31G*(0.25) and CCSD(T)/6-31G*(0.25) energies (Δ(CCSD(T)–MP2)) was added to the MP2/CBS results to give an estimated CCSD(T)/CBS interaction energy. Previous work indicates that the Δ(CCSD(T)–MP2) correction is insensitive to basis set effects,[30] where the 6-31G*(0.25) basis set was previously found to yield satisfactory results for nucleobase or amino acid dimers.[31−35] We note that despite recent literature denoting these extrapolated values as CBS(T) results (see, for example, refs 33, 34, and 36), we will refer to these final interaction energies as CCSD(T)/CBS values to be consistent with our previous work,[21] as well as the majority of previous literature on stacking interactions (see, for example, refs 20, 35, and 37).

All reported interaction energies include basis set superposition error corrections.[38] MP2 and CCSD(T) calculations were performed with Gaussian 03[39] and MolPro,[40] respectively.

Results and Discussion

Dependence of Aromatic−Aromatic Interactions on Structural Characteristics. In contrast to the optimizations typically performed on noncovalently bound dimers, our method for scanning the MP2/6-31G*(0.25) potential energy surface provides more information than simply the optimal stacked and T-shaped orientations. Specifically, our calculations also provide information on how adjusting the relative orientations of monomers in different ways affects the strength of each interaction. Along with our minima structures, this knowledge can be used to understand the relative strength of interactions between nucleobases and amino acid side chains in a variety of orientations. For example, our data can be used to understand contacts observed in nature as demonstrated in the final Results and Discussion section, using aromatic−aromatic DNA−protein interactions found in the PDB.

As a consequence of the large number of data points and dimers considered in this study, select examples will be discussed in this section to illustrate our major findings of the dependence of aromatic−aromatic interactions on structural characteristics. However, all MP2/6-31G*(0.25) binding strengths for all dimers considered are provided in the Supporting Information, which allows the reader to gain a more complete understanding of any given dimer.

For both stacked and T-shaped geometries, we find that the interaction energies are not highly dependent upon changes in the vertical separation between the monomers (R_1). Specifically, for all dimers examined in this study, the interaction energy changes by less than 1.3 kJ mol^{-1} when R_1 deviates from the optimal vertical separation distance by 0.1 Å. Stacked dimers typically have a larger dependence on the vertical separation between monomers than T-shaped dimers, where this trend is generally independent of the monomers involved.

The interaction energy generally has a much larger dependence on the relative angle of rotation (α) than on the vertical separation (R_1). Furthermore, the magnitude of this dependence is related to the type of interaction, as well as the monomers involved. In stacked dimers, α rotations affect the dipole−dipole interactions between monomers, where the dipole moments of the nucleobase and amino acid are antialigned in preferred structures. Therefore, the degree that the interaction energy is dependent on α is largely governed by the magnitude of the dipole moments of the monomers. For instance, since PHE (modeled as benzene) has no dipole moment, the stacking interaction energy has little to no

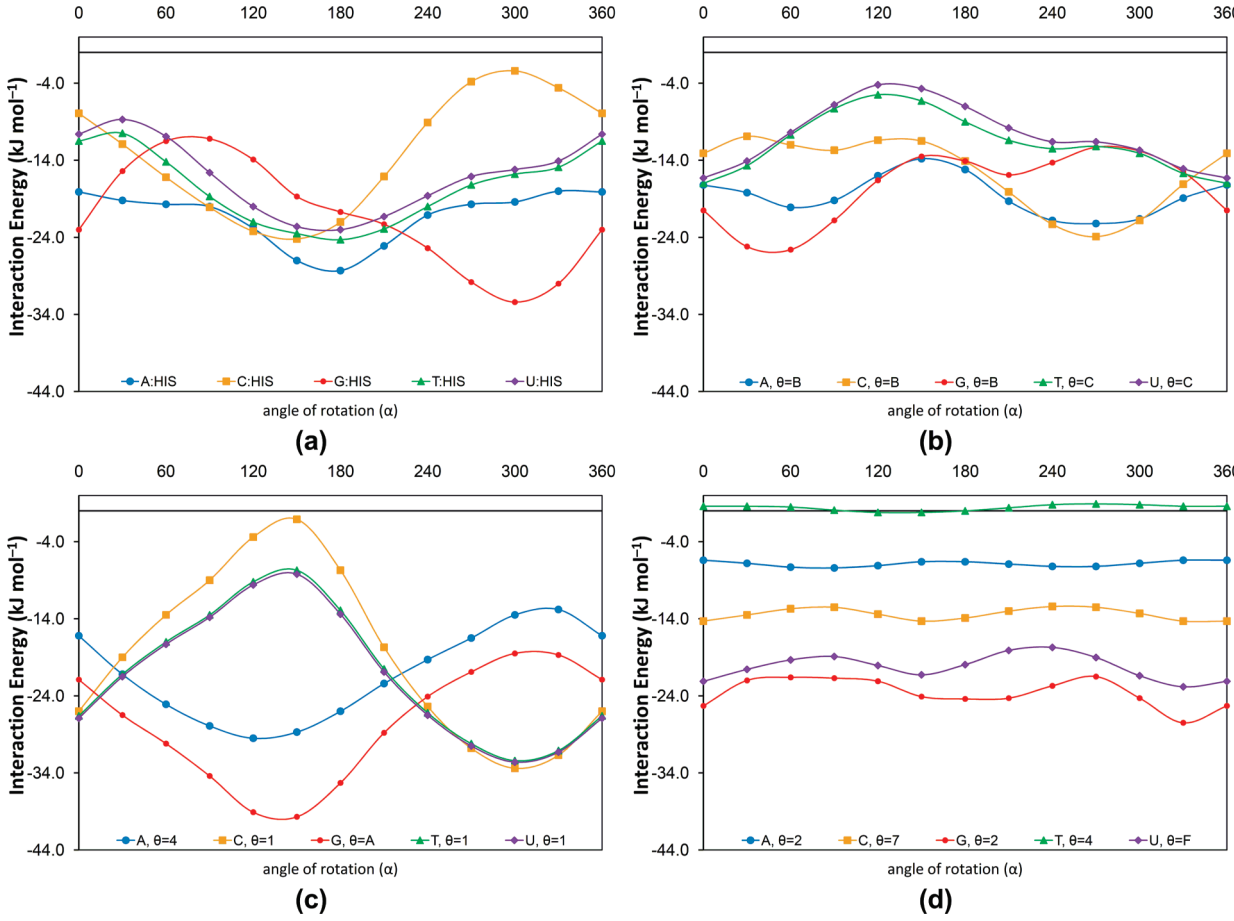


Figure 3. Interaction energy as a function of α rotation for (a) HIS stacked dimers, (b) HIS (edge) T-shaped dimers with the optimal θ, (c) HIS (face) T-shaped dimers with optimal θ, and (d) HIS (face) T-shaped dimers for θ with the weakest α dependence. In all graphs, dimers with adenine are large blue circles, cytosine are orange squares, guanine are small red circles, thymine are green triangles, and uracil are purple diamonds. See Figure 4 for dimer geometries.

dependence on α. However, HIS has the largest dependence on the dipole−dipole alignment (Figures 3a and 4a) since it is the amino acid with the largest dipole moment. Indeed, the α dependency of HIS dimers ranges from 10.3 kJ mol^{-1} for A (which has the smallest nucleobase dipole moment) to 21.8 kJ mol^{-1} for C (which has the largest nucleobase dipole moment).

In T-shaped dimers, the geometric variable α affects how the monomer edge aligns with the monomer face, where the preferred orientation arises when acid/base interactions or secondary intermolecular hydrogen bonds are maximized. For example, in HIS edge interactions, acidic HIS protons align with basic nucleobase sites (see Figure 4b, where the preferred alignment of secondary intermolecular hydrogen bonds are indicated with dotted lines). Therefore, the relative strengths of the intermolecular hydrogen bonds govern the α dependence. Furthermore, since these secondary hydrogen-bonding interactions in amino acid edge dimers are weaker than the dipole−dipole interactions in stacking interactions, the α dependency is smaller in amino acid edge dimers. For example, the α dependence for optimal HIS edge dimers (Figure 3b) ranges from 5.1 (G) to 13.0 (C) kJ mol^{-1}.

Similar to amino acid edge dimers, α rotations in nucleobase edge dimers are affected by the strength of secondary intermolecular hydrogen bonds. For example, Figures 4c and 4d illustrate how nucleobase edges prefer to align basic or acidic edge atoms with the acidic N−H or basic N atom of the HIS face, respectively. When the intermolecular bonds are strong, as found for nucleobase edges with the strongest interactions (see, for example, HIS face dimers in Figure 4c), the α dependence is very large (up to 32.3 kJ mol^{-1} for HIS dimers in Figure 3c). In fact, the α dependence for these nucleobase edge dimers is greater than that observed for stacked dimers. However, other dimers have much weaker intermolecular hydrogen bonds, and therefore a very small dependence on α. For example, the HIS face dimers in Figure 4d lead to a small (1.0 to 6.0 kJ mol^{-1}) dependence on α (Figure 3d).

Only small changes (less than 2.0 kJ mol^{-1} for 98 of the 159 dimers considered) in the interaction energies are observed when one monomer is shifted in the molecular plane relative to the other (R_2 scans) from (0 Å, 0 Å) to the preferred R_2. Additionally, 104 of the 159 R_2 scans reveal that the monomers prefer to be shifted by 1.0 Å or less. These relatively small energetic and geometrical effects indicate that optimal interaction energies are generally observed when the centers of mass of the nucleobase and amino acid are aligned (stacked and amino acid edge interactions) or when the nucleobase edge interacts with the amino acid center of mass (see full computational methods in the Supporting Information). When R_2 shifts strengthen stacked dimers, the preferred shift allows for more $\pi-\pi$ overlap and/or decreases

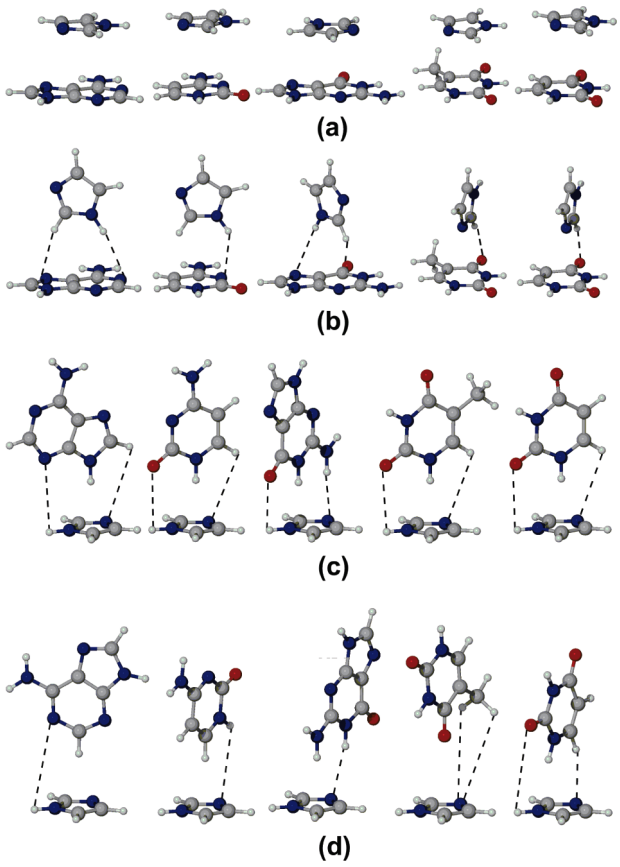


Figure 4. Stacked and T-shaped R_2 = (0 Å, 0 Å) dimers for (left to right) adenine, cytosine, guanine, thymine, and uracil with (a) HIS, (b) HIS (edge) with optimal θ, (c) HIS (face) with optimal θ, and (d) HIS (face) for θ with the weakest α dependence.

repulsion between ring atoms. However, in T-shaped dimers, the preferred R_2 moves the acidic and/or basic regions in the monomer edge closer to the corresponding basic and/or acidic sites in the monomer face to maximize electrostatic interactions.

As previously mentioned, a large number of both amino acid and nucleobase edges (θ) were considered for T-shaped dimers. All T-shaped interactions examined in this study have a very large dependence on the nature of the monomer edge (θ) interacting with the π-system. This large dependence is due to variations in the acid−base properties of different monomer edges, where monomers with large dipole moments (and therefore a large variance in the acid−base properties of its edges) led to a larger dependence on θ. The strongest interactions occur when the monomer edges with the largest (Lewis) acidity are directed toward the π-system. For example, in nucleobase edge dimers, the strongest interactions generally involve the most acidic model (N−H) glycosidic bond of the nucleobases (Figure 4c). However, in the case of guanine, the strongest interaction occurs when two strongly acidic N1−H and N2−H bonds[41] interact with the amino acid face (θ = A, Figure 1) rather than one strongly acidic bond[41] (θ = 5, Figure 1) or one strongly and one weakly acidic bond[41] (θ = E, Figure 1). Similarly, in amino acid edge dimers, the acidic N−H bond of TRP is included in the preferred edge (θ = A), while in HIS edge dimers with A, G, or C (Figure 4b), the optimal T-shaped interaction involves the acidic N−H bond (θ = B, Figure 4b). The only exception is the HIS edge dimer with T or U, where the basic N atom of HIS (θ = C) interacts with the nucleobase π-system in the optimal T-shaped dimer (Figure 4b). This exception occurs since T and U have more positive π-systems and therefore have a greater (Lewis) acidity, due to the two electron-withdrawing carbonyl groups. Recent electron affinity data supports this statement since T and U have the largest electron affinities among all DNA and RNA nucleobases.[42] However, it should also be mentioned that the acidic HIS edge preferred by A, C, and G (θ = B) binds by only 0.4 and 1.1 kJ mol^{-1} weaker to T and U, respectively, than the basic (θ = C) edge.

Since our scans reveal how nucleobase−amino acid aromatic−aromatic contacts depend on several geometric variables, we can draw some general conclusions about the nature of these attractive interactions. First, although stacking interactions are known to be highly dependent on dispersion, it has been shown that additional special $\pi-\pi$ (dispersion) interactions are also significant for π-systems larger than those considered in the present study.[43] We find that our stacking interactions are also highly dependent on the electrostatic contribution. Specifically, our results indicate that the strongest interactions occur when the nucleobase and amino acid dipole moments are anti-aligned, and the interaction energies vary by up to 24 kJ mol^{-1} when one monomer is rotated (by α, Figure 2) relative to the other. Additionally, we find that electrostatics play an even bigger role in T-shaped interactions compared with stacking. For example, the binding strengths of T-shaped dimers vary by up to 56 kJ mol^{-1} by changing the edge (θ, Figure 2) interacting with the π-system. Thus, our potential energy surface scans show that, although dispersion interactions are very important for the attractive nature of these interactions, electrostatics also play an essential role in determining the relative orientations of monomers.

Magnitude of the Strongest Aromatic−Aromatic Interactions. As previously mentioned, monomer orientations with the strongest interactions were further examined by extrapolating the corresponding interaction energies to the CCSD(T) level at the complete basis set (CBS) limit. All stacked dimers, the strongest amino acid edge dimers, the strongest nucleobase edge dimers that involve the model N−H glycosidic bond (relevant for nucleobases), and the strongest nucleobase edge dimers that do not involve the model glycosidic bond (relevant for nucleotides and nucleosides) were examined, where the MP2/6-31G*(0.25) and CCSD(T)/CBS interaction energies are reported in Table 2.

Our calculations reveal that MP2/6-31G*(0.25) interaction energies account for 80−110% of the CCSD(T)/CBS binding strengths in all nucleobase−amino acid dimers (Table 2). Interestingly, the Δ(CCSD(T)–MP2) correction is repulsive for all interaction energies examined, while the Δ(MP2/CBS–MP2/6-31G*(0.25)) correction is attractive. Therefore, our MP2/6-31G*(0.25) interaction energies are very close to the CCSD(T)/CBS results because of the cancelation of errors between basis set incompleteness and the missing electron correlation effects. Nevertheless, the remarkable agreement between these methods verifies that MP2/6-

Table 2. MP2/6-31G*(0.25) and CCSD(T)/CBS Stacked and T-Shaped Interaction Energies (kJ mol^{-1}) between the Aromatic Amino Acids and the Natural DNA or RNA Nucleobases[a]

	A[g]			C			G			T			U		
	θ^b	MP2[c]	CCSD(T)[d]	θ^b	MP2[c]	CCSD(T)[d]	θ^b	MP2[c]	CCSD(T)[d]	θ^b	MP2[c]	CCSD(T)[d]	θ^b	MP2[c]	CCSD(T)[d]
HIS stacked[f]		−29.7	−29.8		−26.9	−29.5		−35.3	−37.8		−25.0	−27.1		−23.6	−26.0
HIS′ stacked[f]		−27.2	−27.8		−26.0	−29.1		−31.4	−32.7		−26.8	−29.3		−26.4	−29.1
PHE stacked[f]		−24.3	−23.3		−18.4	−20.6		−25.3	−25.3		−22.4	−24.5		−20.1	−21.7
TYR stacked[f]		−28.9	−28.4		−24.2	−27.1		−32.8	−35.2		−26.1	−28.3		−25.2	−27.6
TYR′ stacked[f]		−30.7	−30.0		−22.7	−25.0		−33.4	−34.1		−25.5	−27.5		−25.7	−27.7
TRP stacked[f]		−32.0	−30.6		−33.4	−35.4		−42.4	−42.1		−36.0	−37.8		−34.5	−36.8
TRP′ stacked[f]		−35.0	−33.5		−32.9	−35.4		−42.5	−43.1		−36.4	−38.5		−34.0	−36.1
HIS (edge)	B	−22.5	−22.4	B	−26.4	−27.5	B	−27.7	−27.2	C	−19.1	−23.3	C	−18.4	−22.8
PHE (edge)	A	−14.1	−15.2	A	−12.1	−14.0	A	−14.5	−16.0	A	−10.8	−12.3	A	−9.7	−11.7
TYR (edge)	A	−21.9	−22.4	A	−22.5	−23.7	A	−21.8	−21.8	A	−15.7	−16.6	1	−14.9	−18.0
TRP (edge)	A	−23.2	−22.9	A	−27.4	−28.8	A	−28.7	−28.9	A	−20.9	−22.1	A	−18.3	−19.2
HIS (face)[e]	4 (8)	−33.6 (−22.6)	−33.3 (−23.6)	1 (4)	−33.6 (−31.2)	−36.1 (−33.7)	5 (A)	−26.7 (−46.4)	−26.7 (−48.4)	1 (3)	−34.2 (−22.5)	−34.2 (−24.4)	1 (3)	−34.7 (−23.0)	−36.9 (−24.9)
PHE (face)[e]	4 (8)	−25.6 (−16.0)	−26.4 (−17.4)	1 (5)	−22.5 (−19.5)	−25.4 (−21.8)	5 (A)	−23.7 (−31.6)	−24.3 (−33.2)	1 (3)	−24.7 (−18.9)	−22.2 (−21.6)	1 (3)	−25.3 (−19.1)	−27.6 (−21.7)
TYR (face)[e]	4 (8)	−27.8 (−18.4)	−28.3 (−20.0)	1 (4)	−26.7 (−25.3)	−29.1 (−28.2)	E (A)	−24.6 (−36.2)	−24.9 (−37.8)	1 (3)	−27.6 (−20.6)	−29.4 (−22.9)	1 (3)	−27.8 (−20.8)	−29.7 (−23.1)
TRP (face)[e]	4 (8)	−34.8 (−23.1)	−34.8 (−24.0)	1 (4)	−34.5 (−32.2)	−36.6 (−33.9)	E (A)	−32.5 (−47.3)	−32.4 (−49.5)	F (3)	−37.0 (−25.4)	−37.9 (−27.2)	F (3)	−36.7 (−25.7)	−37.7 (−27.4)

[a] See full computational methods provided in the Supporting Information. [b] See Figures 1 and 2 for definitions of θ and the edges considered in this study. [c] MP2/6-31G*(0.25) interaction energies. [d] CCSD(T)/CBS interaction energies. [e] Nucleobase edge interactions involving the model glycosidic bond and not involving the glycosidic bond (in parentheses). [f] Reference 18. [g] Reference 21.

31G*(0.25) very accurately describes the potential energy surfaces of nucleobase−amino acid dimers in both stacked and T-shaped orientations and suggests that we are correctly identifying the preferred orientations for these aromatic−aromatic contacts. This also suggests that the MP2/6-31G*(0.25) interaction energies reported in the Supporting Information for all relative monomer orientations are reliable and can be used to characterize interactions observed in nature (to be discussed in a subsequent section).

One of our most important findings is the magnitude of these aromatic−aromatic interactions. At the most accurate level of theory used to date for nucleobase−aromatic amino acid stacked and T-shaped dimers (CCSD(T)/CBS), we find that stacking interactions range between −20 and −43 kJ mol^{-1}, while the corresponding T-shaped interactions range between −12 and −50 kJ mol^{-1} at optimal monomer orientations. Thus, both the stacking and T-shaped interactions approach the adenine–thymine Watson−Crick hydrogen-bond strength (−70 kJ mol^{-1}) calculated at the same level of theory,[35] which involves at least two strong hydrogen bonds. Additionally, we find that the binding strengths for stacking and T-shaped interactions are very similar in magnitude for all nucleobases, where the amino acid edge interactions are slightly weaker than the nucleobase edge interactions. Indeed, the most favorable nucleobase edge interactions are as strong, if not stronger, than the corresponding stacking interactions in the same dimer. Therefore, stacking arrangements are not the only orientations that are significantly attractive, but rather T-shaped interactions can also play stabilizing roles in nucleobase−amino acid complexes found in nature.

Comparison of the largest interaction energy for each dimer reveals interesting information about the trends in the binding energy with respect to the amino acid or nucleobase. Specifically, the strongest stacking interactions are found to increase as PHE ≪ TYR ≈ HIS < TRP, where similar stabilization energies are observed for HIS, TYR, and TRP. This trend is dictated by the relative dipole moments and size of the π-system of the amino acids. The same trend is also found for amino acid edge and nucleobase edge T-shaped interactions, which depend on the strength of secondary intermolecular hydrogen bonds between the acidic and basic sites in the amino acid and the nucleobase. However, when the trend with respect to the nucleobase is considered, no clear affinity is observed. The implications of these findings will be further discussed in the next section.

The Role of DNA−Protein Aromatic−Aromatic Interactions in Nature. A previous statistical survey of DNA−protein and RNA−protein interactions suggests that some aromatic amino acids could be involved in nucleic acid bases recognition because of a large number of observed contacts (Table 1).[15] For example, HIS forms the greatest number of contacts with G and T and therefore may selectively bind to these bases, while PHE forms the greatest number of contacts with T and A. Conversely, it was concluded that TYR provides strength to both DNA and RNA−protein complexes rather than having recognition capabilities since there is no clear affinity for one nucleobase.

Even though stacking and T-shaped interactions between the nucleobases and aromatic amino acids are strong, they can only play a role in DNA (RNA) recognition if there are large deviations in the binding energies with a change in the nucleobase. However, for a specific amino acid, we find that the differences between the binding strengths of all nucleobases are less than 10.5 kJ mol^{-1}. Additionally, in the nucleobase trend for a given amino acid, the difference in the binding strengths for any two sequential bases is generally very small (less than 3 kJ mol^{-1} for 82% of the cases). Therefore, our results indicate that none of the aromatic amino acids preferentially bind to any of the nucleobases through stacking or T-shaped contacts. Furthermore, comparison of our calculated binding strengths (Table 2) to the number of observed contacts (Table 1) indicates that there is no significant correlation between the frequency of a particular nucleobase−amino acid contact and the magnitude of the interaction energy of the dimer.

We note that the lack of significant correlation between the frequency of a particular nucleobase−amino acid contact (Table 1)[15] and our calculated interaction energies may arise for many reasons. First, the original study of Baker and Grant[15] searched the PDB for DNA−protein structures involving only double stranded DNA, where many other relevant contacts may not fit this restriction. Second, the observed frequency of these interactions in the PDB may not reflect the natural occurrence of these contacts since crystal structures are not available for many DNA−protein complexes because of, for example, difficulties crystallizing native forms of the complexes. Third, classifying each aromatic−aromatic interaction as a specific type of contact is not easy and therefore determining their overall frequency and importance is difficult. Additionally, our results in Table 2 report the magnitude of these interactions in the preferred stacked or T-shaped orientations, where the geometries (and therefore magnitude) in PDB structures may be quite different. We also only considered dimers and interactions with additional residues may affect the magnitude of the binding interaction. Finally, since biomolecular interactions are so complex, the strongest interaction is not always the interaction that leads to optimal biological function. Therefore, there may not be a simple justification for expecting a correlation between the frequency of a particular nucleobase−amino acid contact and our strongest interaction energies (see, for example, comments in refs 36, 44, and 45).

All of the factors discussed above likely have a large effect on the correlation between the frequency and magnitude of DNA/RNA−protein interactions. Therefore, it is difficult to determine whether aromatic−aromatic interactions are selective enough to be used for recognition of specific DNA nucleobases. Nevertheless, our calculations at the highest level of theory possible for these systems demonstrate that aromatic−aromatic stacking and T-shaped interactions can be very stabilizing. Therefore, our calculations suggest that these interactions are most certainly important for providing stability to DNA/RNA−protein complexes.

Additional T-Shaped Contacts That May Stabilize DNA/RNA−Protein Complexes. As previously mentioned, our method for examining nucleobase−amino acid T-shaped interactions reveals more information than simply the identity of edges that lead to the strongest interactions. Specifically, we characterized the potential energy surfaces of all atom-directed and bridge-directed edges for these systems, where our full results (Supporting Information) show that many monomer edges lead to attractive interactions. For instance, we characterized a large number of C−H$\cdots\pi$ interactions, which have been extensively studied computationally,[46−52] and found that these nucleobase−amino acid interactions can be up to −26 kJ mol^{-1}. For example, in TRP edge dimers, the bridged edge involving two C−H$\cdots\pi$ interactions (θ = C) is only 0.8 (U) to 7.8 (A) kJ mol^{-1} weaker than the strongest interactions observed with a bridged edge involving a N−H and C−H bond (θ = A). Additionally, in TRP face dimers, interactions involving only one C−H$\cdots\pi$ contact are up to −18.5 (A, θ = 5) or −18.9 (G, θ = 6) kJ mol^{-1}.

We also examined lone pair (lp)$\cdots\pi$ interactions for a variety of dimers, which have been studied computationally[50,53] and recently reviewed by Egli et al.[54] In nucleobase edge dimers, lp$\cdots\pi$ interactions involving carbonyl oxygens were found to be unstable (i.e., no minima was identified for U (θ = 2 and θ = 4) and C (θ = 2)) or very weak (i.e., less than −1 kJ mol^{-1} for G (θ = 8) and T (θ = 2 and θ = 4)). However, some nitrogen lp$\cdots\pi$ interactions are as strong as −23 kJ mol^{-1}. For example, all HIS edge dimers involving the lone pair N (θ = 3) were found to be quite stable, where the interaction energies range from −8.2 to −15.2 kJ mol^{-1}. Furthermore, the interactions with N7 of G (θ = 7) can be up to −13.8 kJ mol^{-1}, while contacts with N3 of C (θ = 3) were found to be up to −23.0 kJ mol^{-1}. Therefore, our calculations suggest that even nonoptimal contacts that might arise in nature due to, for example, protein folding, can help stabilize DNA/RNA−protein complexes.

Quantifying Experimentally Observed Nucleobase−Amino Acid Aromatic−Aromatic Interactions. We acknowledge that the most favorable structures and interaction energies of nucleobase−amino acid dimers identified in our calculations may not appear in nature, where additional effects, such as protein folding or other (steric) constraints, come into play. However, by carefully searching the potential energy surface of stacked and T-shaped dimers, we have revealed how the interaction energies in these noncovalent complexes generally depend on different geometric variables. For example, because of the small dependence on the vertical separation between monomers (R_1), deviations from our reported optimal R_1 distances will have a small effect on the interaction energies. Conversely, large deviations from our reported preferred angle of rotation (α) between stacked monomers will have significant effects on the interaction energy between monomers with large dipole moments. Alternatively, deviations from our optimal α for T-shaped dimers will affect the interaction energy in different ways, and therefore the properties of the edge involved must be carefully considered. Our calculations also suggest that changes in the optimal horizontal displacement (R_2) between monomers can weaken the (strongest) reported interaction energies, where the effect is dependent on the monomers involved, as well as the size and direction of the shift.

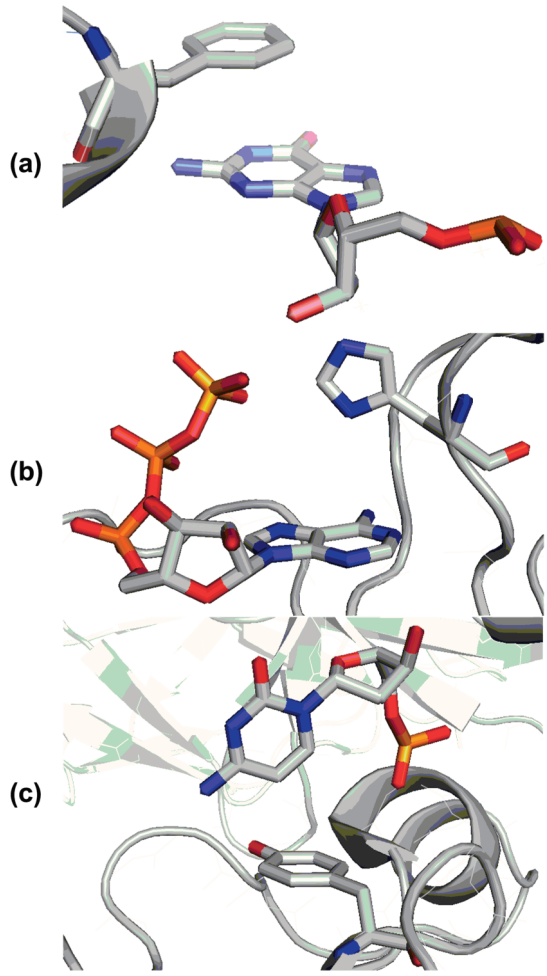


Figure 5. Examples of nucleobase−amino acid interactions observed in nature (a) G:PHE stacked (PDB code: 1ckt with DG109 and F37 highlighted)[55] (b) A(face):HIS(edge) T-shaped (PDB code: 1b8a with ATP1500 and H1223 highlighted),[56] and (c) C(edge):TYR(face) T-shaped (PDB code: 1p7h with DC11 (chain B) and Y424 (chain M) highlighted).[57]

In addition to the general conclusions mentioned above, we can use the results from our detailed potential energy surface scans in a quantitative, predictive way. Specifically, in addition to the optimal interaction energies presented in Table 2, the interaction energies calculated for a range of geometric variables are reported in the Supporting Information for all nucleobase−amino acid dimers. These results can be compared to any stacked or T-shaped contact found in nature, not just those corresponding to minima on our calculated surfaces. This complete data set will allow biochemists to determine the relative magnitude and importance of different DNA/RNA−protein contacts.

To illustrate how our data can be used in conjunction with experimental structures to gain a better understanding of DNA−protein interactions that appear in nature, we will discuss three crystal structures containing representative examples of nucleobase−amino acid stacking (Figure 5a),[55] amino acid edge T-shaped (Figure 5b),[56] and nucleobase edge T-shaped (Figure 5c)[57] interactions. In this section, our Supporting Information will be used to estimate the interaction energy of each dimer, which will be compared to the interaction energy calculated using the exact crystal structure orientation (see crystal structure overlay section in the computational methods of the Supporting Information for details of how the interaction energies were calculated in crystal structure orientations).

Stacking Interaction between Guanine and Phenylalanine. The crystal structure of HMG1 domain A bound to a cisplatin-modified DNA duplex contains a stacked orientation between guanine (DG109) and phenylalanine (F37).[55] Table 3 reports relevant geometric parameters measured from the crystal structure that define this stacking orientation (column "Crystal"). These parameters can subsequently be compared to the closest corresponding point on our scanned potential energy surface (column "Scan"). For example, in this crystal structure, the orientation between G and PHE is most closely described by the point on our calculated potential energy surface with a vertical separation (R_1) of 3.5 Å, an angle of rotation (α) of 30°, and a horizontal displacement (R_2) of (1.5 Å, 1.5 Å). Our calculations (Table SI-2, guanine–phenylalanine) reveal that an R_2 shift of (1.5 Å, 1.5 Å) has a corresponding interaction energy of −22.9 kJ mol^{-1}. However, this interaction was calculated for a (R_1) vertical separation distance of 3.4 Å and $\alpha = 0°$. Nevertheless, corrections based on our scans can be added to this interaction energy. Specifically, our R_1 scan reveals that increasing the vertical separation from 3.4 to 3.5 Å weakens the interaction energy by 0.6 kJ mol^{-1}. Additionally, an α rotation from 0 to 30° weakens the interaction energy by 0.1 kJ mol^{-1}. Therefore, the interaction between G and PHE in the crystal structure geometry of HMG1 domain A is estimated from our potential energy surface scans to be −22.2 kJ mol^{-1} (i.e., (−22.9 + 0.6 + 0.1) kJ mol^{-1}). Interestingly, when optimized monomers are overlaid onto this crystal structure, the calculated interaction energy is also found to be −22.2 kJ mol^{-1}. This is in astounding agreement with our estimate from the calculated potential energy surfaces, especially since we have assumed additivity of the effects of different geometric variables.

Amino Acid Edge Interaction between Histidine and Adenine. In the crystal structure of aspartyl-tRNA synthetase, the edge of histidine (H1223) interacts with the face of adenine (ATP1500).[56] Since protein crystal structures are not able to determine the positions of hydrogen atoms, it is not clear whether the HIS N atom or N−H bond is directed toward A. Therefore, both HIS edge interactions ($\theta = 3$ and $\theta = 1$, Figure 2) are considered, where Table 3 outlines the geometric information obtained from the crystal structure and identifies the closest corresponding point on our potential energy surface scan for each histidine orientation ($R_1 = 4.5$ Å, $R_2 = (0.5$ Å, 2.5 Å), and $\alpha = 60°$ ($\theta = 3$) or 300° ($\theta = 1$), see Table SI-2, adenine–histidine).

When the nitrogen lone pair of HIS is directed toward A ($\theta = 3$), the interaction energy is calculated to be −6.2 kJ mol^{-1} for the α value (60°) closest to that in the crystal structure, $R_1 = 4.4$ Å and $R_2 = (0$ Å, 0 Å). To determine the binding in the crystal structure orientation, our potential energy surface scans reveal that the binding strength is weakened by 8.1 kJ mol^{-1} when HIS is shifted across the face of A from (0 Å, 0 Å) to (0.5 Å, 2.5 Å), and weakened

Table 3. Structural Characteristics Measured for Three Representative Nucleobase−Amino Acid Interactions Observed in Nature (Crystal) and the Closest Corresponding Point from Our Potential Energy Surface Scans (Scan)

	G:PHE stacked[51]		A(face):HIS(edge)[52]		C(edge):TYR(face)[53]	
	crystal	scan	crystal	scan	crystal	scan
angle (°)	4.44[a]	0	88.90[a]	90	82.72[a]	90
vertical separation distance (R_1, Å)	3.51[b]	3.5	4.48[b]	4.5	4.02[c]	3.0[d]
angle of rotation (α, °)	18.8[e]	30	67.8[e]	60 ($\theta = 3$), 300 ($\theta = 1$)	66.3[e]	60, 300
horizontal shift distance (R_2, Å and (Å,Å))	∼2.2[f]	(1.5, 1.5)	∼3.1[f]	(0.5, 2.5)	∼2.7[f]	(2.5, 0)

[a] Angle measured between molecular planes formed by aromatic ring atoms of nucleobase and amino acid. [b] Vertical separation measured between the molecular planes (stacked orientation), or the center of the ring (monomer edge) and monomer face molecular plane (T-shaped orientations). [c] Vertical separation measured between C5 of C and molecular plane of TYR. [d] The R_1 distance from our potential energy surface scans measured between the midpoint of the H4−C5 edge of C (θ = E) and the center of mass of TYR, which has the closest corresponding vertical separation measured between C5 of C and the center of the TYR ring. See Supporting Information for computational methods and definitions of R_1 used in this study. [e] The dihedral angle measured between the glycosidic bond and the peptide backbone bond of the amino acid (stacked orientation), the glycosidic bond and C1′ plane and the molecular plane of the amino acid (amino acid edge), or the peptide backbone bond plane of the amino acid and the molecular plane of the nucleobase (nucleobase edge). [f] The distance estimated between the centers of monomer rings in two dimensions.

by 0.1 kJ mol^{-1} when the vertical separation is increased to 4.5 Å. These additive effects result in a weakly (+2 kJ mol^{-1}) repulsive interaction for the $\theta = 3$ HIS(edge):A(face) interaction, while crystal structure overlay calculations determine this interaction to be only slightly more favorable (−2.1 kJ mol^{-1}).

When the N−H bond of HIS is directed toward A ($\theta = 1$, Figure 1), our potential energy surface scan determines the interaction energy to be −17.3 kJ mol^{-1} for the angle of rotation ($\alpha = 300°$) and vertical separation distance ($R_1 = 4.5$ Å) observed in the crystal structure. Although this interaction energy was calculated for $R_2 = (0$ Å, 0 Å), accounting for the correct horizontal displacement yields an interaction energy of −9.2 kJ mol^{-1}. In comparison, the crystal structure overlay calculations determine this interaction to be −10.0 kJ mol^{-1}. This example again provides strong evidence that our calculated potential energy surfaces accurately describe the interaction energies in a range of nucleobase−amino acid dimers.

Nucleobase Edge Interaction between Cytosine and Tyrosine. The crystal structure of NFAT1 bound to the HIV-1 LTR kB element reveals the edge of C (DC11 (chain B)) interacting with the face of TYR (Y424 (chain M).[57] Because of an uncertainty in the orientation of the hydroxyl hydrogen of TYR, we consider both possible (in-plane) orientations, namely OH directed toward and away from C. The nucleobase edge interaction in the crystal structure with the hydroxyl group directed away from C (θ = E) best corresponds to $\alpha = 60°$, which has a calculated interaction energy of −18.8 kJ mol^{-1} for $R_1 = 2.6$ Å and $R_2 = (0$ Å, 0 Å) (Table SI-2, cytosine–tyrosine). When a (∼10 kJ mol^{-1}) correction for the horizontal shift and a (3.2 kJ mol^{-1}) correction for the vertical separation are taken into account, the interaction energy is predicted to be −5.6 kJ mol^{-1}, which is very consistent with the interaction energy calculated from a crystal structure overlay (−4.7 kJ mol^{-1}). Alternatively, when the hydroxyl group of TYR is directed toward C (θ = E, $\alpha = 300°$), the interaction energy is estimated to be −3.5 kJ mol^{-1} (i.e., −16.7 ($\alpha = 300°$, $R_1 = 2.6$ Å, $R_2 = (0$ Å, 0 Å)) + 3.2 (R_1 correction) + 10.0 kJ mol^{-1} (R_2 correction)), while our overlay calculation determines an interaction energy of −5.5 kJ mol^{-1}.Therefore, our predictions using the data set generated from our potential energy surface scans are very similar to those determined from crystal structure overlay calculations.

In summary, this section illustrates how our data reported in the Supporting Information can help predict the stabilizing and functional role of nucleobase−amino acid aromatic−aromatic interactions in a variety of biological structures. Our results can therefore be used in the future by biochemists in a similar fashion to accurately estimate the magnitude of these interactions in DNA/RNA−protein active sites identified through experimental structure determination. Thus, our data complements experimental structural data and helps reveal the relative importance of aromatic−aromatic contacts in a variety of DNA/RNA−protein systems.

Conclusion

This is the first extensive study of both stacking and T-shaped interactions between all aromatic amino acids and natural nucleobases using our unique methodology for identifying important contacts on the potential energy surface. The variables used in this study allow us to understand how a variety of geometric criteria determine the strength of these contacts. The interaction energies were generally found to have a small dependence on the vertical separation between the monomers, while the dependence on the angle of rotation, as well as horizontal displacement, was found to be much greater. Additionally, stacked complexes prefer to have monomer dipole moments anti-aligned, while preferred T-shaped orientations direct the most (Lewis) acidic edge of one monomer toward the π-system of the second monomer.

Our ab initio calculations reveal that the strongest possible aromatic−aromatic interaction for each nucleobase−amino acid combination is very large. Indeed, at the highest level of theory possible for these systems, we find that both stacking and T-shaped interactions are very close in magnitude to biologically relevant hydrogen bonds. Additionally, T-shaped interactions are as strong, if not stronger, than the corresponding stacking interactions. Most importantly, we have not only fully characterized structures with the strongest interactions, but we have investigated a wide variety of other nucleobase−amino acid interactions, including C−H$\cdots\pi$, N−H$\cdots\pi$, and lp$\cdots\pi$ interactions, and our calculations reveal that these contacts are also very favorable.

On the basis of our calculations, we stress that most aromatic−aromatic nucleobase−amino acid interactions are large enough to play important roles in biological processes and therefore cannot be ignored. However, we do not find a correlation between the maximum strength of a specific interaction and its frequency in nature, nor do we find a large difference between the binding strengths with the natural bases for any given amino acid. Therefore, it is difficult to determine whether these interactions play a role in selective nucleobase recognition. Nevertheless, because of the large magnitude of their π-interactions, we suggest that the aromatic amino acids can take advantage of their π-clouds to add to the stability of DNA/RNA−protein complexes already provided by hydrogen-bonding interactions with these or other protein side chains.

Our high accuracy results (CCSD(T)/CBS) indicate that the MP2/6-31G*(0.25) binding strengths provided in the Supporting Information are very precise for all 485 dimers investigated in this study. Therefore, the MP2 results can provide reliable estimates of the size of interactions present in a variety of experimental DNA/RNA−protein crystal structures. Indeed, we have demonstrated how this additional, detailed data set can be used to estimate dimer interaction energies for contacts found in the Protein Data Bank, where our estimates are very close to interaction energies calculated using exact crystal structure orientations. Thus, this manuscript illustrates how our results can be used by biochemists to accurately approximate the magnitude and relative importance of π-interactions in DNA/RNA−protein actives sites identified in their experiments.

Although our results correspond to contacts within binding sites of low polarity,[58] solvation effects will be important to consider when extending our work to highly polar environments, such as those found in common thermodynamic experiments, where electrostatic forces between monomers are diminished.[59] Therefore, future work will consider environmental effects on the magnitude of these interactions, which will also extend the applicability of our results to all DNA/RNA−protein binding sites. Future work will also analyze the strength of specific contacts found in a greater number of crystal structures, as well as the effects of more than one simultaneous interaction on the noncovalent contacts between the nucleobases and aromatic amino acids.

Acknowledgment. We thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Research Chair Program, the Canadian Foundation for Innovation (CFI), and the University of Lethbridge Research Fund for financial support. L.R.R acknowledges NSERC, Alberta Ingenuity Fund (AIF), the Alberta Government, and the University of Lethbridge for student scholarships.

Supporting Information Available: Full computational methods, MP2/6-31G*(0.25) interaction energies for all points considered in potential energy surface scans, optimal R_1, α, and R_2 values for each θ examined, and all high-level interaction energies required to estimate CCSD(T)/CBS binding strengths. Coordinates for all dimers in Table 2 are available from the authors upon request. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Gromiha, M. M.; Siebers, J. G.; Selvaraj, S.; Kono, H.; Sarai, A. *Gene* **2005**, *364*, 108–113.

(2) Chan, L. L.; Pineda, M.; Heeres, J. T.; Hergenrother, P. J.; Cunningham, B. T. *ACS Chem. Biol.* **2008**, *3*, 437–448.

(3) Höglund, A.; Kohlbacher, O. *Proteome Sci.* **2004**, *2*, 3.

(4) Ptashne, M. *Nature (London)* **1967**, *214*, 232–234.

(5) Bartsevich, V. V.; Miller, J. C.; Case, C. C.; Pabo, C. O. *Stem Cells* **2003**, *21*, 632–637.

(6) Luscombe, N. M.; Thornton, J. M. *J. Mol. Biol.* **2002**, *320*, 991–1009.

(7) Nadassy, K.; Wodak, S. J.; Janin, J. *Biochemistry* **1999**, *38*, 1999–2017.

(8) Pailard, G.; Lavery, R. *Structure* **2004**, *12*, 113–122.

(9) Pabo, C. O.; Nekludova, L. *J. Mol. Biol.* **2000**, *301*, 597–624.

(10) Matthews, B. W. *Nature (London)* **1988**, *335*, 294–295.

(11) Hogan, M. E.; Austin, R. H. *Nature (London)* **1987**, *329*, 263–266.

(12) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11163–11168.

(13) Luscome, N. M.; Laskowski, R. A.; Thornton, J. M. *Nucleic Acids Res.* **2001**, *29*, 2860–2874.

(14) Lejeune, D.; Delsaux, N.; Charloteaux, B.; Thomas, A.; Brasser, R. *Proteins: Struct. Funct. Bioinf.* **2005**, *61*, 258–271.

(15) Baker, C. M.; Grant, G. H. *Biopolymers* **2007**, *85*, 456–470.

(16) Cauët, E.; Rooman, M.; Wintjens, R.; Lievin, J.; Biot, C. *J. Chem. Theory Comput.* **2005**, *1*, 472–483.

(17) Cysewski, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2636–2645.

(18) Rutledge, L. R.; Campbell-Verduyn, L. S.; Wetmore, S. D. *Chem. Phys. Lett.* **2007**, *444*, 167–175.

(19) Mao, L.; Wang, Y.; Liu, Y.; Hu, X. *J. Mol. Biol.* **2004**, *336*, 787–807.

(20) Copeland, K. L.; Anderson, J. A.; Farley, A. R.; Cox, J. R.; Tschumper, G. S. *J. Phys. Chem. B* **2008**, *112*, 14291–14295.

(21) Rutledge, L. R.; Wetmore, S. D. *J. Chem. Theory Comput.* **2008**, *4*, 1768–1780.

(22) Biot, C.; Buisine, E.; Rooman, M. *J. Am. Chem. Soc.* **2003**, *125*, 13988–13994.

(23) Hobza, P.; Šponer, J. *Chem. Rev.* **1999**, *99*, 3247–3276.

(24) Šponer, J.; Leszczynski, J.; Hobza, P. *J. Mol. Struct. (THEOCHEM)* **2001**, *573*, 43–53.

(25) Šponer, J.; Leszczynski, J.; Hobza, P. *Biopolymers* **2002**, *61*, 3–31.

(26) Hobza, P. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2004**, *100*, 3–27.

(27) Rutledge, L. R.; Durst, H. F.; Wetmore, S. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2801–2812.

(28) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(29) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437–446.

(30) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.

(31) Hobza, P.; Šponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.

(32) Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613.

(33) Jurečka, P.; Šponer, J.; Hobza, P. *J. Phys. Chem. B* **2004**, *108*, 5466–5471.

(34) Šponer, J.; Jurečka, P.; Marchan, I.; Javier Luque, F.; Orozco, M.; Hobza, P. *Chem.—Eur. J.* **2006**, *12*, 2854–2865.

(35) Jurečka, P.; Šponer, J.; Černý, J; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(36) Šponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595–2610.

(37) Pitoňák, M.; Neogrády, P.; Řezáč, J.; Jurečka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829–1834.

(38) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(39) 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 D.02; Gaussian Inc., Wallingford, CT, 2004.

(40) 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, Version 2006.1; University College Cardiff Consultants Ltd., Cardiff, UK, 2006.

(41) McConnell, T. L.; Wheaton, C. A.; Hunter, K. C.; Wetmore, S. D. *J. Phys. Chem. A* **2005**, *109*, 6351–6362.

(42) Roca-Sanjuán, D.; Merchán, M.; Serrano-Andrés, L.; Rubio, M. *J. Chem. Phys.* **2008**, *129*, 095104.

(43) Grimme, S. *Angew. Chem., Int. Ed.* **2008**, *47*, 3430–3434.

(44) Kopitz, H.; Živković, A.; Engels, J. W.; Gohlke, H. *ChemBioChem* **2008**, *9*, 2619–2622.

(45) Luo, R.; Gilson, H. S. R.; Potter, M. J.; Gilson, M. K. *Biophys. J.* **2001**, *80*, 140–148.

(46) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 3746–3753.

(47) Ringer, A. L; Figgs, M. S.; Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10822–10828.

(48) Shibasaki, K.; Fujii, A.; Mikami, N.; Tsuzuki, S. *J. Phys. Chem. A* **2006**, *110*, 4397–4404.

(49) Gil, A.; Branchadell, V.; Bertran, J.; Oliva, A. *J. Phys. Chem. B* **2007**, *111*, 9372–9379.

(50) Mishra, B. K.; Sathyamurthy, N. *J. Phys. Chem. A* **2007**, *111*, 2139–2147.

(51) Tsuzuki, S.; Fujii, A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2584–2594.

(52) Tsuzuki, S.; Honda, K.; Fujii, A.; Uchimaru, T.; Mikami, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2860–2865.

(53) Tsuzuki, S.; Mikami, M.; Yamada, S. *J. Am. Chem. Soc.* **2007**, *129*, 8656–8662.

(54) Egli, M.; Sarkhel, S. *Acc. Chem. Res.* **2007**, *40*, 195–205.

(55) Ohndorf, U. M.; Rould, M. A.; He, Q.; Pabo, C. O.; Lippard, S. J. *Nature (London)* **1999**, *399*, 708–712. PDB Code: 1ckt.

(56) Schmitt, E.; Moulinier, L.; Fujiwara, S.; Imanaka, T.; Thierry, J. C.; Moras, D. *EMBO J.* **1998**, *17*, 5227–5237. PDB Code: 1b8a.

(57) Giffin, M. J.; Stroud, J. C.; Bates, D. L.; von Koenig, K. D.; Hardin, J.; Chen, L. *Nat. Struct. Biol.* **2003**, *10*, 800–806. PDB Code: 1p7h.

(58) Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5291–5303.

(59) Florián, J.; Šponer, J.; Warshel, A. *J. Phys. Chem. B* **1999**, *103*, 884–892.

CT800567Q

JCTC Journal of Chemical Theory and Computation

Reduced Catalytic Activity of P450 2A6 Mutants with Coumarin: A Computational Investigation

Weihua Li,[†,‡] Hirotaka Ode,[†] Tyuji Hoshino,*[,†] Hong Liu,[§] Yun Tang,*[,‡] and Hualiang Jiang[‡,§]

Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 263-8522, Japan, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China, and Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

Received January 11, 2009

Abstract: Human cytochrome P450 2A6 is the major enzyme to catalyze coumarin 7-hydroxylation, and this enzyme also plays an important role in the metabolism of nicotine and other tobacco-specific compounds. Recent experimental data showed that the N297S and A481T mutants of P450 2A6 decreased the catalytic activity toward coumarin by about 4-fold and 10-fold, respectively. These two mutants also had about 30-fold decrease in binding affinity for coumarin when compared to its wild type. At present, however, how the mutations affect the enzymatic activity and/or the substrate binding remains unclear. In this study, a combination of molecular docking and molecular dynamics (MD) simulation was employed to investigate the above question. Our results demonstrated that the N297S mutation altered the hydrogen-bonding network mediated by a water molecule between the B′−C loop and the I helix and thus a shift of the B′ helix/B′−C loop region, whereas the A481T mutation triggered the conformational changes of its adjacent residues including Phe209 and Phe280 via an indirect manner to affect the substrate binding. However, the mutations did not significantly alter the substrate binding orientation because the only polar residue 297 in the active site provided the hydrogen-bonding donor to guide the binding of coumarin. Both mutations perturbed the shape of "Phe-cluster" in the active site and thus weakened the interactions with coumarin. The calculated binding free energies were in agreement with the relative potency of the experimental binding affinities.

1. Introduction

Cytochromes P450 (CYPs) are a superfamily of heme-containing monooxygenases. These enzymes play key roles not only in the metabolism and detoxification of a large number of xenobiotic chemicals but also in the biosynthesis of many endogenous compounds.[1] Cytochrome P450 2A6 (CYP2A6) is one of the 57 CYP isoenzymes found in humans, which exists mainly in the liver and is also expressed in extrahepatic tissues.[2]

CYP2A6 can specifically catalyze coumarin 7-hydroxylation,[3] which has been used as an isoform-specific marker reaction for estimating individual's CYP2A6 activity in vivo.[4] Another important aspect of CYP2A6 is its role in the metabolism and activation of some carcinogenic compounds. For example, it was shown to be one of the major enzymes responsible for nicotine metabolism.[5] In addition to nicotine, CYP2A6 also catalyzes activation of many tobacco-specific nitrosamines and other toxic compounds, including NNN (*N*′-nitrosonicotine),[7] NNK (4-(methylnitrosamino)-1-(3-pyridy)-1-butanone),[6,7] and MTBE (methyl *t*-butyl ether).[8] For this reason, CYP2A6 has been proposed to be a potential drug target for

* Corresponding author phone: +81-43-2902926; fax: +81-43-2902925; e-mail: hoshi@p.chiba-u.ac.jp (T.H.), phone: +86-21-64251052; fax: +86-21-64253651; e-mail: ytang234@ecust.edu.cn (Y.T.).

[†] Chiba University.

[‡] East China University of Science and Technology.

[§] Chinese Academy of Sciences.

10.1021/ct900018t CCC: $40.75
Published on Web 03/30/2009

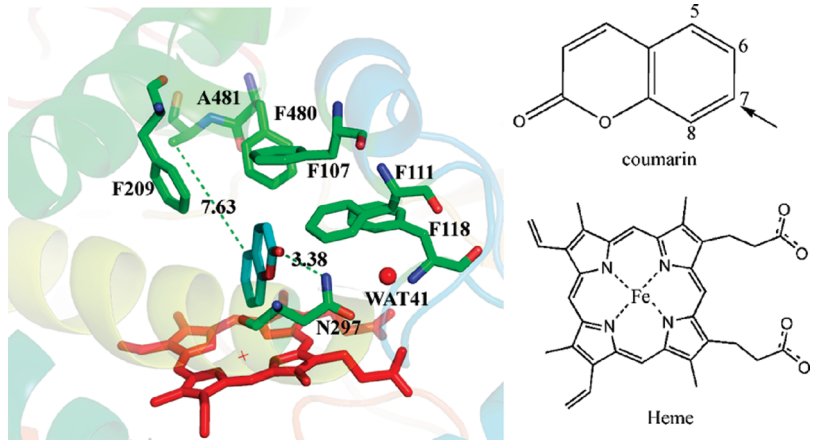


Figure 1. A close view of coumarin bound in the active site of CYP2A6 (PDB code 1Z10). Asn297, Ala481 and the Phe-cluster are labeled. The distances (Å) between coumain and Asn297 as well as Ala481 are labeled. The water molecule mediated between N297 and the B′−C loop is shown in red sphere. The 2D structures of coumarin and heme are shown on the right side. The arrow indicates the oxidation site of coumarin by CYP2A6.

smoking cessation.[9] Like other CYPs, CYP2A6 also contributes to the metabolism of several drugs, such as tegafur[10] and losigamone.[11] In addition, CYP2A6 is a genetically polymorphic enzyme.[12] More than 30 CYP2A6 variant alleles have been identified so far (http://www.imm.ki.se/CYPalleles/cyp2a6.htm).

Furthermore, CYP2A6 has indicated its potential application in the pharmaceutical industry. Work from Guengerich group demonstrated that CYP2A6 was a good biocatalyst to catalyze the oxidation of indoles to indigo and indirubin.[13] Subsequent studies from the same group found that some CYP2A6 mutants were more active than its wild type (WT) in forming indigo.[14] Several CYP2A6 mutant enzymes generated with a randomized library approach showed high inhibitory activity against human cyclin-dependent kinases (CDK)-1 and -5 and glycogen synthase kinase-3 (GSK3), two important drug targets related to a variety of diseases, for instance, cancer.[15]

Recently, Kim et al.[16] discovered several CYP2A6 mutants exhibiting altered catalytic activities with coumarin by using a combination technique of random mutagenesis and high-throughput screening. Their results demonstrated that CYP2A6 N297S and A481T mutants decreased catalytic efficiency (K_{cat}/K_m) toward coumarin by 4 and 10 times or so, respectively. These two mutants also had about 30-fold decrease in binding affinity for coumarin when compared to WT. They suggested that the reduced binding affinity was a major basis for the decreased catalytic efficiency.[16] The crystal structure of WT CYP2A6 in complex with coumarin was determined recently,[17] which provides useful static structural information on the active site. The structure reveals that coumarin is confined within the active site consisting of a "Phe-cluster" formed by residues Phe107, Phe111, Phe108, Phe209, and Phe480, as shown in Figure 1. Residue Asn297 functions as a hydrogen-bonding donor to interact with coumarin directly (Figure 1). Accordingly, it was generally believed that replacement of Asn297 by other residues might change the direct interaction between the residue 297 and coumarin.[16,18] However, when Asn297 is substituted with serine, the direct hydrogen-bonding interaction may still exist because the hydroxyl group of serine is also able to serve as a hydrogen-bonding donor. Hence, how the N297S mutation decreases the binding affinity of CYP2A6 with coumarin is poorly understood. As for Ala481, the crystal structure shows that this residue is about 7.6 Å away from the substrate and does not directly interact with the substrate (Figure 1). Therefore, it also remains unclear how the A481T mutation affects the substrate binding.

In the present work, to elucidate the mechanism underlying the reduced catalytic activity of CYP2A6 for coumarin due to residual mutation, molecular docking combined with molecular dynamic simulations and binding free-energy calculations was performed on the WT and mutant systems. Our results indicate that both N297S and A481T mutations trigger the conformation changes of the Phe-cluster and thus alter the shape of the active site of the enzymes. These alterations weaken the interactions of the substrate with the mutated enzymes. The calculated binding free energies are in agreement with the relative potency of coumarin bound with CYP2A6 WT and the mutants.

2. Methods

2.1. Construction of N297S and A481T Models. To date, several CYP2A6 crystal structures including different ligand-free mutants and ligand-bound complexes have been determined, but there is only one crystal structure available for CYP2A6 in complex with coumarin (PDB code 1Z10, at 1.9 Å resolution).[17] Therefore, for consistency, the protein models of the substrate-free N297S and A481T mutants were constructed based on the available 2A6 WT-coumarin crystal structure. The mutations at positions 297 and 481 were accomplished by *Sybyl7.0* (Tripos Inc.). The resulting mutant N297S and A481T models were subjected to energy-minimization. The detailed protocol for the minimization was described in the subsequent MD simulations section. After the minimization, the mutant models were refined by 1.0 ns MD simulations with position restraints on the protein backbone atoms excluding the residues within 5.0 Å around the mutated residues.

2.2. Molecular Docking. The substrate, coumarin, was docked into the active sites of the MD-refined CYP2A6

N297S and A481T mutants using *GOLD* version 3.2.[19] The Chemscore scoring function with specific parameters[20] for heme-containing proteins was used to rank the docking poses. This scoring function has been validated to outperform the Goldscore scoring function when applied to CYP docking by previous studies.[20,21] The Fe atom of the heme group was chosen as the center for docking and the residues within 20.0 Å around this atom were defined as the binding site. Twenty solutions were output for each docking and ranked according to Chemscore. The pose with lowest Chemscore was selected as the initial conformation for the further long-time MD simulation.

2.3. Parameterization of Heme−Thiolate Complex and Coumarin. To derive appropriate force field parameters for heme group used for molecular dynamics simulations, quantum chemical calculation was performed on the heme−thiolate complex. According to the generalized catalytic cycle of P450[22] and the crystal structure of CYP2A6-coumarin,[17] binding of the substrate displaced the water molecule that is coordinated to iron and led to a low to high spin conversion of heme. Therefore, we model the heme−thiolate complex in its pentacoordinate ferric form with the sextet spin state. The fifth coordination of iron was treated as S−Cys, and the acetyl (ACE) and *N*-methyl (NME) groups as the capping residues for Cys439. The geometrical optimization of heme-thiolate complex was performed at HF/6-31G* level. Subsequently, population analysis was carried out with the B3LYP method.[23,24] For the Fe atom, the triple-ζ-valence (TZV) basis set was used, and for H, C, N, O, and S atoms, the 6-31G* basis set was employed. On the basis of the calculated electron population, atomic charges were derived using restrained electrostatic potential (RESP) fitting procedure,[25] which ensures that the derived atom charges are compatible with the standard AMBER force field. For the force constant parameters involving Fe, we adopted the values that were kindly provided by Dr. Harris from his previous work.[26]

Atom charges of coumarin were obtained using the similar procedure as above. Geometrical optimization and the electrostatic potential were executed at the B3LYP/6-31G** level. The RESP procedure implemented in the *Amber9* antechamber module was utilized to derive the partial atomic charges of coumarin. All quantum chemical calculations were executed using the *Gaussian03* program.[27]

2.4. Protein Preparation for MD Simulations. The initial structural model of CYP2A6 WT complexed with coumarin for MD simulation was taken from the Protein Data Bank (PDB entry code 1Z10).[17] The PDB file of 1Z10 contains 4 asymmetric 2A6 molecules. Molecule A was used to present the results of the modeling and MD simulations in this study. For verification of the results, molecule B also was repeated to perform the same modeling and simulations as molecule A. Therefore, a total of 12 MD simulations were performed in this study. Six MD simulations (3 position-restrained simulations for substrate-free systems of WT, N297S and A481T, each by 1 ns; 3 unrestrained simulations for substrate-bound complexes of WT, N297S and A481T, each by 6 ns) were done for molecule A of 1Z10. Likewise, 6 MD simulations were carried out for molecule B of 1Z10 for the purpose of validation of our conclusion. It should be noted that Leu370 in molecule A has a split conformation. For consistency with molecule B, the conformation with the side chain pointing to coumarin was used for Leu370 in this study. The protonation states of charged residues and histidines were determined on the basis of pK_a values calculated by *PROPKA*.[28] According to the calculation results, His229 and His254 were assigned to be fully protonated at both nitrogen atoms. His72 and His477 were assigned to be protonated at ε nitrogen and other histidines to be protonated at δ nitrogen atoms. In addition, Glu448 was set in the protonated state (pK_a = 8.24).

An all-atom model of CYP2A6 WT was generated using the xleap module in *Amber9*.[29] All crystallization water molecules were kept in the initial model. The initial models of 2A6 N297S and A481T mutants complexed with coumarin for MD simulations were obtained from the docking results. Because only a few residues in the active site underwent the conformational changes in the two mutant models, the coordinates of all crystallization water molecules were copied into the mutant models from 1Z10 to keep as consistent with WT as possible in the initial states. The resulted models were then solvated with water molecules in a truncated octahedron periodic box. The distance between the box walls and the protein was set to 10.0 Å, which resulted in about 13 000 water molecules for each system. The TIP3P[30] water model was used. The systems were neutralized by adding the corresponding number of counterions.

2.5. MD Simulations. All of the MD simulations were performed using the *Amber9* package.[29] The AMBER ff03 all atom force field[31] was used for the protein and the general AMBER force field[32] was used as the parameters for coumarin. Energy minimization was conducted in three steps. First, only water molecules and ions were allowed to move. Next, the movement was extended to the residues within 5.0 Å around the mutated residue and the substrate (if any). Finally, all atoms were allowed to move freely. In each step, energy minimization was carried out by a combination of the steepest descent method for 5000 steps and the conjugated gradient method for another 5000 steps. After the minimization, each system was gradually heated from 0 to 300 K over 60 ps under the NVT ensemble condition and equilibrated at 300 K for 100 ps. Finally, 6.0 ns unrestrained MD simulations were conducted at 1 atm and 300 K under the NPT ensemble condition. During the simulation, the SHAKE algorithm[33] was applied to constrain the covalent bonds to hydrogen atoms. A time step of 2.0 fs and a nonbonding interaction cutoff radius of 10.0 Å were used. Coordinates were saved every 1.0 ps during the entire process.

2.6. Calculation of Volume of the Active Site. The volume of the active site of CYP2A6 as well as its N297S and A481T mutants was estimated using the *Proshape Pocket* program.[34] The program is based on the Alpha Shape theory, which provides an analytic method for detecting pockets in proteins and measuring their volume and surface area. A probe radius of 1.4 Å was employed. Because the program cannot recognize the heme group, we subtracted the volume of heme group (approximately 200 Å^3) from the preliminary calculation results as suggested by the previous work.[35]

2.7. Binding Free Energy Calculations. The binding free energy between coumarin and the protein was calculated with the MM-PBSA approach,[36−38] according to the following equation:

$$\Delta G_{\text{binding}} = \Delta G_{\text{MM}} + \Delta G_{\text{solv}} - T\Delta S \quad (1)$$

where $\Delta G_{\text{binding}}$ is the binding free energy, ΔG_{MM} is the molecular mechanical energy, ΔG_{solv} is the solvation energy, and $T\Delta S$ is the entropy contribution. The molecular mechanical energy is calculated by the following equation:

$$\Delta G_{\text{MM}} = \Delta G_{\text{int}} + \Delta G_{\text{elec}} + \Delta G_{\text{vdw}} \quad (2)$$

where ΔG_{int}, ΔG_{elec}, and ΔG_{vdw} represent internal, electrostatic, and van der Waals energy in the gas phase, respectively. The solvation energy is divided into two components:

$$\Delta G_{\text{solv}} = \Delta G_{\text{ele,sol}} + \Delta G_{\text{nonpol,sol}} \quad (3)$$

where $\Delta G_{\text{ele,sol}}$ is the electrostatic contribution to solvation energy, and $\Delta G_{\text{nonpol,sol}}$ is the nonpolar solvation term. Here, the polar contribution was calculated by solving the Poisson−Boltzmann equation, whereas the latter is determined using,

$$\Delta G_{\text{nonpol,sol}} = \gamma(\text{SASA}) + b \quad (4)$$

where γ represents surface tension and b is constant, whereas SASA is the solvent-accessible surface area (Å^2).

In this study, 200 snapshots from the last 2.0 ns of production stage were extracted for binding free energy calculations. The polar contribution term of solvation energy was calculated using the *PBSA* program in *Amber9*. The interior dielectric constant was set to 1.0, and the outer dielectric constant was set to 80.0. The solvent-accessible surface area was determined using the LCPO method.[39] The coefficient γ and b were set to 0.0072 kcal /(mol·Å^2) and 0 respectively, as in the work of Still and co-workers.[40] Normal mode analysis[41] was conducted to estimate the entropic changes using the nmode program in *Amber9*. Because the current CYP2A6 systems are relatively large (about 8000 atoms excluding water and ions) and thus very memory demanding, only residues within 12.0 Å of the substrate were included for the normal mode calculations. This treatment has been used in many previous studies.[26,42] The truncated systems were minimized for up to 100 000 cycles to give an energy gradient of 0.0001 kcal·mol^{-1}·Å^{-1} using a distance-dependent dielectric constant of $\varepsilon = 4r$. Fifty snapshots of each system were selected for the entropy calculation.

3. Results and Discussion

At present, only the crystal structure of 2A6 WT in complex with coumarin has been determined, and no structural models are available for the N297S and A481T mutant complexes. Therefore, first we need to construct the mutant complexes models. The crystal structure of WT−coumarin complex[17] reveals that the substrate forms a direct hydrogen bond with Asn297, which implies that the replacement of Asn297 by other residues could change the binding orientation of the substrate in the active site of the enzyme. In addition, previous studies on other CYPs indicated that a single residual mutation could indeed result in different binding orientations of the ligand in the active site of the enzymes.[43,44] Bearing these facts in mind, we adopted the following protocol to investigate the effect of two mutations on the catalytic activity of CYP2A6. First, we constructed the protein models of substrate-free mutants based on the crystal structure of WT−coumarin complex. Then, the initial complex models of mutants were built by docking coumarin into the active site of substrate-free mutant models. After that, long-time MD simulations of the WT and mutant complexes were performed to refine the substrate binding modes and explore the dynamic behaviors of proteins and the substrate in the active site of the enzymes. Finally, binding free energy was calculated to analyze the energy differences between WT and the mutants.

3.1. Models of N297S and A481T Mutant Proteins. The protein models of substrate-free N297S and A481T mutants were obtained by a 1.0 ns MD simulation and then energy-minimization. For comparison, the substrate-free WT was also built using the same procedure as the mutant construction. The root-mean-square deviations (rmsd) of backbone atoms of N297S and A481T with respect to WT were 0.20 Å and 0.18 Å, respectively. Comparison of the mutant models with WT revealed that only a few residues adjacent to the mutated residues underwent slight displacement, primarily for the phenylalanine residues in the Phe-cluster. In the N297S mutant, Ser297 formed a stable hydrogen bond via the side-chain hydroxyl group with the backbone oxygen atom of its neighboring residue Met293. The phenyl ring of Phe111 in the N297S mutant had a slight rotation compared to WT. The same event has been observed in the crystal structures of other substrate-free CYP2A6 mutants including N297Q, N297Q/L241C, and N297Q/I300V.[18] All of these mutants exhibit no significant backbone changes but slight displacement in side chains of certain residues in the active sites. For the A481T mutant, the side chains of Phe209 and Phe480 underwent a slight displacement when compared to WT. Both residues are situated around the mutated Thr481.

3.2. Complex Models of N297S and A481T with Coumarin by Docking. The initial complex models of N297S and A481T with coumarin were obtained by using the docking program *GOLD*. Prior to obtaining the mutant complexes using *GOLD*, the binding mode of WT with coumarin was reproduced by docking the substrate into the MD-refined substrate-free WT model. The docking result showed that coumarin displayed similar binding orientation to that of the crystal structure of WT−coumarin except a slight translation of the planar ring of coumarin. This is not surprising because coumarin is planar and small in size. For consistency, the comparison of the mutant complexes with respect to WT was based on the docked WT−coumarin complex model.

Compared with the WT−coumarin complex, the binding mode of coumarin in the active site of N297S differed significantly (Figure S1 of the Supporting Information). Because Ser297 formed a hydrogen bond with Met293, no hydrogen bond between Ser297 and coumarin was observed.

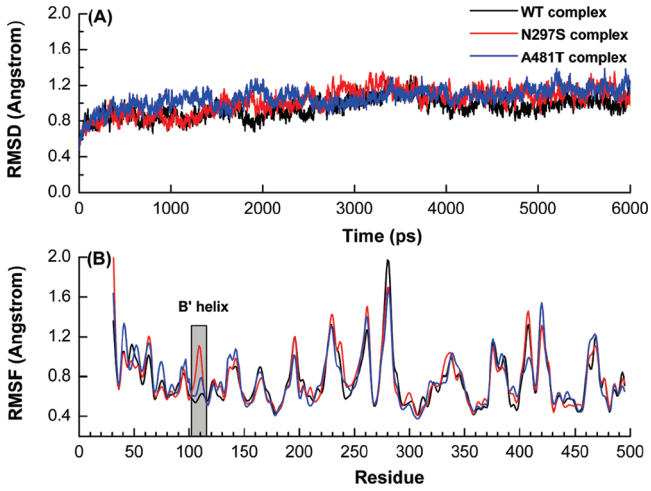


Figure 2. (A) rmsd variations of CYP2A6 WT (black), N297S (red), and A481T (blue) mutants with respect to the starting structures as a function of simulation time. (B) RMSF of non-hydrogen atoms versus residue position of WT (black), the N297S mutant (red), and the A481T mutant (blue) calculated from MD simulations. The shadow region represents the B′ helix.

Although coumarin kept a binding pose similar to that in the WT complex, the planar ring had an obvious deviation in the N297S complex. Because the hydrogen bond is abolished and thus the altered binding mode, the Chemscore of coumarin with N297S (27.4) was lower than that of coumarin with WT (29.2). In contrast, the binding mode of coumarin in the A481T mutant was almost the same as that of the WT complex. A direct hydrogen bond was formed between Asn297 and coumarin in the A418T complex. Because Asn297 is the only polar residue in the active site of CYP2A6 and the side chain is pointing to the active site, this residue easily guides the binding orientation of the substrates that have hydrogen-bonding acceptors. The docking score was 28.9 for coumarin with A481T. Neither the substrate binding modes nor the docking scores can differentiate the interactions between WT and A481T with coumarin. Several recent studies[45,46] clearly demonstrated that automated docking alone was not suitable for accurate discrimination of the ligands and/or proteins that have structurally high similarity. However, docking combined with MD simulations was shown to successfully achieve the goal. Therefore, the 2A6 WT and the mutant complexes were subsequently subjected to 6 ns unrestrained MD simulations.

3.3. MD Simulations of 2A6 Complexes. MD simulations were performed on each complex with two purposes. The first is to refine the substrate binding modes because the docking does not take into account the flexibility of proteins. The other purpose is to explore the dynamic behavior of the proteins and the substrate in the active site of the enzymes during the long-time MD simulations.

Overall Structural Changes. The rmsd values of protein backbone atoms with the individual starting structure were monitored along the whole MD process. Part A of Figure 2 shows the rmsd of three systems as they vary with respect to simulation time. The rmsd values of three complexes have a large fluctuation during the first 4.0 ns and reach stability after 4.0 ns. The protein atoms do not undergo significant structural changes and the rmsd values of the three systems converge to about 1.1 Å, a relatively small deviation from the minimized structure. Because the rmsd values showed no significant fluctuation after 4.0 ns in all three systems, subsequent analysis was performed on the snapshots from the last 2.0 ns.

To examine the mobility of protein residues, the root-mean-square fluctuation (RMSF) of backbone atoms of each residue was investigated, as shown in part B of Figure 2. The three structures have similar profiles in RMSF distribution except for the shadow region, which represents the B′ helix. This region displays relatively large RMSF values in the N297S mutant. This mainly attributes to the N297S mutation, which altered the interactions between this region and the I helix, as discussed in the following section. Apart from the N-terminus, residues with the large RMSF value belong to the loops connected with two well-conserved secondary structures, α helices, or β sheets. Because these loops are located at the surface of the protein and completely exposed to the solvent, it is not surprising that these regions have a large RMSF. This trend of RMSF calculated by our simulations is also in accordance with deduction from the crystal structure.[17]

Conformational Change of Residues in the Active Site. To address the structural difference between WT and the mutants during the long-time MD simulations, we analyzed the conformational changes of residues in the active sites. The average structures of three complexes were first calculated with 200 snapshots, which were extracted from the last 2.0 ns. Then, the average structures of the N297S and A481T mutants were superimposed on that of WT with protein backbone atoms. The superposition demonstrates that these three protein structures are similar in whole. The rmsd value of the N297S mutant with WT is 0.84 Å and it is 0.95 Å between the A481T mutant and WT. However, detailed analysis of the individual residues at the active site and comparison of local structures provide new insights into the structural changes due to mutations.

Comparison of the N297S mutant with WT revealed a movement of the B′−C loop involving residues Val116-Ser119, and a slight outward shift of the B′ helix occurred in the N297S mutant, as shown in part A of Figure 3. Several Phe residues of the Phe-cluster in the N297S mutant, such as Phe107, Phe111, and Phe118, underwent displacement when compared to WT. The shift of the B′ helix/B′−C loop region resulted in an expansion of the active site and thus an enlargement in the volume of the active site of the N297S mutant. To confirm this view, we estimated the volumes of active sites of WT and the N297S mutant. As expected, the average volumes of those of WT and the N297S mutant are 147 Å^3 and 167 Å^3, respectively. The active site volume of the N297S mutant is about 14% larger than that of WT. The calculated volumes are comparable to the values calculated by another group using the same program.[35]

Detailed analysis of the MD trajectories of N297S and WT indicated that the shift of the B′ helix/B′−C loop region was primarily caused by the change of the hydrogen-bonding network around residue 297. The crystal structure of WT complexed with coumarin[17] reveals that Asn297 interacts

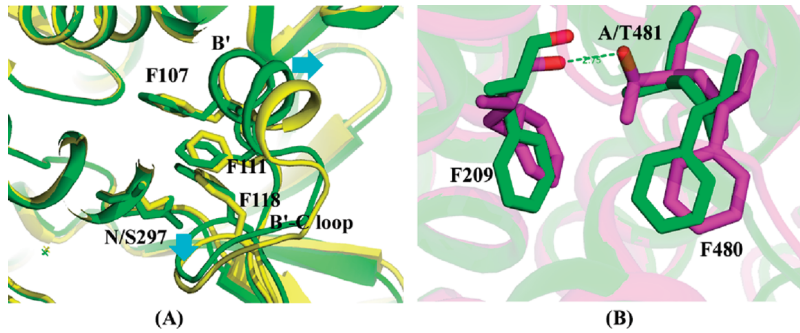


Figure 3. Superposition of average structures from the last 2 ns of the three complexes. (A) Superposition between WT (green) and the N297S mutant (yellow). (B) Superposition between WT (green) and the A481T mutant (magenta). Important secondary structure elements and the residues with significant structural changes are labeled. The arrow indicates the movement of the secondary elements.

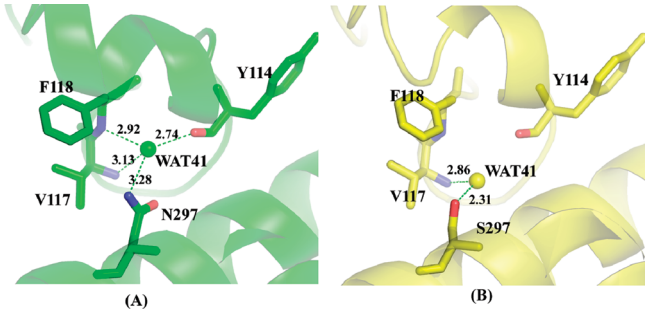


Figure 4. Hydrogen bonding network mediated by WAT41 in MD snapshots of (A) the WT complex and (B) the N297S complex. Hydrogen bonds are represented by the green dotted lines and the distances (Å) between heavy atoms are labeled.

with the B′−C loop via a water molecule (WAT41) mediation. The water molecule forms hydrogen bonds with the backbone N atoms of both Val117 and Phe118, the backbone O atom of Tyr114, and the side chain N atom of Asn297. During the MD simulation of the WT complex, the water molecule kept the hydrogen-bonding network interactions the same as those of the crystal structure. WAT41 formed hydrogen bonds with Asn297, Val117, Phe118, and Tyr114, as shown in part A of Figure 4. The occupancy of these hydrogen bonds is greater than 50% of the last 2.0 ns MD simulation, as shown in Table 1. In contrast, the hydrogen-bonding network mediated by WAT41 in the N297S mutant was changed due to the shortened side chain of Ser297. In the N297S mutant, WAT41 maintained hydrogen-bonding interactions with more than 50% occupancy only with Val117 and Ser297 (part B of Figure 4). The hydrogen-bonding network formed by WAT41 in the A481T mutant was similar to that of WT (Table 1). Because the B′ helix/B′−C loop region directly interacts with the I helix and constitutes a part of the substrate binding pocket, the region was considered to play an essential role in the substrate binding.[18] Previous mutagenesis experiments confirmed that residues in the B′ helix/B′−C loop region affected the catalytic activity of CYP2B1.[47]

In addition to the shift of the B′ helix/B′−C loop region, the phenyl ring of Phe111 in the N297S mutant underwent a rotation compared to WT. To identify the ring rotation, the change of the side-chain torsion χ_2 (CA, CB, CG, and CD1) of Phe111 was monitored for the WT and N297S complexes, as shown in part A of Figure 5. Compared to WT, the torsion angle of Phe111 in the N297S mutant has an obvious change. Phe111 is located above the residue 297 and has hydrophobic interactions with residue 297. The substitution of Asn by Ser at position 297 leads to a shorter side chain and thus changes the interaction with Phe111. This may be the reason for rotation of the Phe111 ring. The rotation of the phenyl ring of Phe111 due to the change of residue 297 has been observed in the crystal structures of CYP2A6 and in other mutants including N297Q and N297Q/L240C.[18]

The crystal structure of 2A6 bound with coumarin[17] reveals that Ala481 is a distance of about 7.6 Å from the substrate and does not directly interact with coumarin. Accordingly, the A481T mutation may influence the substrate binding through an indirect manner, that is, the mutation affects its neighboring residues to transfer the effect. This

Table 1. Hydrogen Bonding Networks Mediated by WAT41 During the Last 2 ns of MD Simualtions

Donor		Acceptor		Occupancy (%)
		WT		
Phe118	H	O	WAT41	97.3
Val117	H	O	WAT41	86.5
Asn297	HD21	O	WAT41	75.5
WAT41	H1	O	Tyr114	58.2
WAT41	H2	O	Tyr114	53.4
WAT41	H1	N	Phe118	24.1
WAT41	H2	N	Phe118	19.9
WAT41	H1	ND2	Asn297	17.8
WAT41	H2	ND2	Asn297	15.6
		N297S		
Val117	H	O	WAT41	99.8
WAT41	H1	OG	Ser297	50.5
WAT41	H2	OG	Ser297	50.1
WAT41	H1	O	Tyr114	26.6
WAT41	H2	O	Tyr114	24.7
		A481T		
Phe118	H	O	WAT41	96.7
Val117	H	O	WAT41	77.5
Asn297	HD21	O	WAT41	74.5
WAT41	H1	O	Tyr114	57.0
WAT41	H2	O	Tyr114	57.6
WAT41	H1	N	Phe118	18.1
WAT41	H2	N	Phe118	16.2
WAT41	H1	ND2	Asn297	16.7
WAT41	H2	ND2	Asn297	15.3

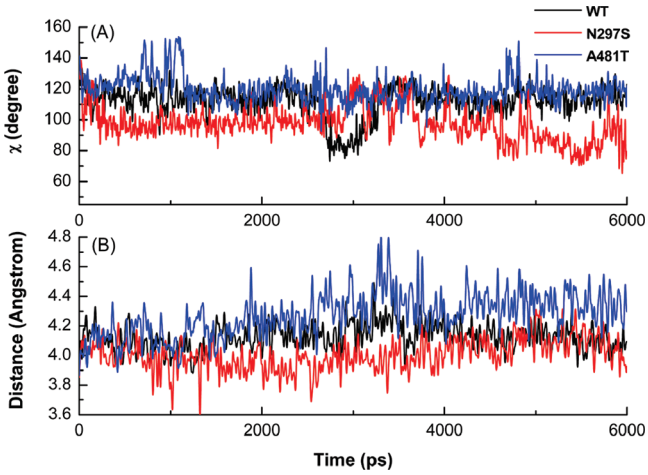


Figure 5. (A) Variation of the side-chain torsion χ_2 (CA, CB, CG, and CD1) of Phe111 in three complexes versus time. (B) The variation of distances between Fe and C-7 of coumarin versus simulation time.

idea was supported by our modeling and MD simulations. Superimposition of WT with the A481T mutant indicated the mutation caused significant displacement of Phe480 and an altered orientation of Phe209 when compared with WT, as shown in part B of Figure 3. Both residues belong to the Phe-cluster mention earlier.

Although the replacement of Ala481 by Thr increases the length and size of residue 481, Thr481 still cannot directly make contact with the substrate. Nevertheless, the mutated Thr481 side chain was enough to trigger the structural changes of its adjacent residues. To accommodate the increased size of residue 481, the phenyl ring of Phe480 moved away from Thr481 by about 1.5 Å relative to WT. On the other hand, the hydroxyl group of the side chain of Thr481 formed a hydrogen bond with the backbone oxygen atom of Phe209. The direct interaction makes the contact between these two residues more intimate. This may be one of the reasons that the orientation of Phe209 was changed in the A481T mutant. Both Phe480 and Phe209 have been demonstrated to substantially alter the catalytic activity of 2A6[14] and other 2A enzymes[48] when the Phe residue at positions 480 or 209 was replaced by other residues. Because both affected residues, Phe480 and Phe209, move away from the active site, the synchronous effect of these residues' movement resulted in the expansion of the volume of the active site in the A481T mutant. The calculated average volume of the active site of the A481T mutant is approximately 177 Å^3, roughly a 20% increase from that of WT.

Substrate Binding Mode. The influence of mutation on the binding modes of coumarin in the active site was investigated

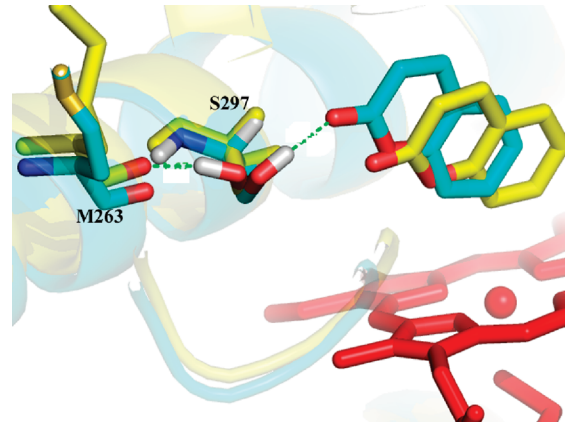


Figure 6. Substrate induced the side-chain conformational change of Ser297 in the N297S mutant during MD simulation. Yellow, before MD simulation; cyan, after MD refinement.

by analyzing the MD trajectories. Superimposition of the average structures of the substrate demonstrated that neither mutation significantly altered the substrate binding modes. Although the active-site volume of the A481T mutant is about 20% larger than that of WT, the alteration did not induce the change in the substrate binding. This mainly contributes to the facts that the active site of CYP2A6 is highly hydrophobic and tightly packed and that only polar residue 297 in the active site provides a hydrogen-bonding donor so as to orient coumarin binding.

Interestingly, the substrate induced a rotation of the side chain of Ser297 during the MD simulation of the N297S mutant. Before MD simulation, the hydroxyl group of Ser297 pointed to Met293 and formed a hydrogen bond with the backbone oxygen atom of Met293. After MD refinement, however, the hydroxyl group pointed to coumarin and formed a hydrogen bond with the oxygen atom of the substrate, as shown in Figure 6. The induced-fit phenomenon caused by ligand binding has been observed in the crystal structures of several P450s, notably CYP2C9 and CYP2B4, for instance. The trivial difference of binding modes between N297S and WT is in agreement with the experimental data. Previous studies indicated that the N297S mutation did not affect the K_{cat} value but only decreased the K_m by about 4-fold for coumarin 7-hydroxylation.[16]

Although similar substrate binding orientations were observed for the A481T mutant and WT, the distances of C-7 of coumarin from the iron atom differed in these two enzymes. Part B of Figure 5 shows the distance variation of C-7 and Fe along the MD simulations of three systems. The distance of C-7 and Fe in the N297S mutant is almost same as that of WT. In contrast, the average distance of C-7 and Fe in the A481T mutant is 0.3 Å longer than that of WT. It

Table 2. Binding Free Energies of P450 2A6 WT, N297S, and A481T Complexed with Coumarin[a]

2A6	ΔG_{ele}	ΔG_{vdw}	$\Delta G_{nonp,sol}$	$\Delta G_{ele,sol}$	$-T\Delta S$	$\Delta G^{cal}_{binding}$	$\Delta G^{exp}_{binding}$ [b]	$\Delta G^{exp}_{binding}$ [c]
WT	−12.20 ± 0.12	−25.45 ± 0.09	−2.90 ± 0.01	20.40 ± 0.08	12.37 ± 0.54	−7.78	−8.19	−7.80
N297S	−14.36 ± 0.13	−23.53 ± 0.10	−3.03 ± 0.01	23.13 ± 0.06	11.76 ± 0.56	−6.03	−6.19	−7.02
A481T	−11.75 ± 0.07	−24.78 ± 0.09	−3.05 ± 0.01	21.45 ± 0.08	12.31 ± 0.56	−5.82	−6.17	−6.99

[a] All energies are in kcal/mol. The uncertainties are the standard error of the mean calculated with 200 snapshots (50 snapshots for entropic calculations). [b] $\Delta G^{exp}_{binding}$: calculated from the experimental data via $\Delta G^{exp}_{binding} \approx RT \ln K$ at $T = 300$ K. The binding energies are calculated based on K_s. [c] The binding energies are calculated based on K_m.

is well-known that the distance between the oxidation distance and Fe can influence the activation energy of the proton and electron transfer reaction.[49] The increase of the distance may lead to a decrease in the reaction rate. This may account for the experimental result that the K_{cat} value of A481T decreases ~3-fold toward coumarin 7-hydroxylation as compared with WT. A previous study also reported a similar finding for CYP2A6 V117A mutation with coumarin 7-hydroxylation.[50]

3.4. Binding Free Energy Analysis. Binding free energies of WT and the mutants were calculated and analyzed with the MM-PBSA method. Table 2 lists the calculated energies, including the total binding energies and the individual energy components. According to Table 2, the WT complex has the lowest binding energy. Binding energies of both mutants are higher than that of WT. The recent biochemical experiment indicated that the K_s and K_m values of both mutants with coumarin were higher than that of WT.[16] Generally, the greater the K_s or K_m value is, the lower binding affinity the enzyme has for substrates. It is encouraging that our predicted binding energies are in good agreement with the relative potency of the experimental binding affinities.

A detailed energy analysis showed that van der Waals energies are the major driving force for the substrate−enzyme binding. This is in line with the fact that the active site of CYP2A6 is mainly composed of hydrophobic residues. Compared to WT, the van der Waals energy shows a considerable increase in the N297S mutant. This was mainly caused by the shift of the B′ helix/B′−C loop region due to the N297S mutation. The electrostatic interaction energy of N297S was more favorable for the substrate binding than that of WT. However, the solvation electrostatic energy and the van der Waals energy canceled the favorable contribution. For the A481T mutant, the difference was not so significant because the mutation affected the substrate binding via an indirect manner. However, the accumulation of the energy differences made the substrate binding unfavorable with the A481T mutant.

4. Conclusions

Molecular docking combined with MD simulations and binding energy calculations was performed on human CYP2A6 as well as its N297S and A481T mutants complexed with coumarin for the purpose of elucidating the mechanisms underlying the reduced catalytic activity and binding affinity due to residual mutations. Our simulations demonstrated that the N297S mutation caused an altered hydrogen-bonding network mediated by a water molecule between the B′−C loop and the I helix and thus a shift of the B′ helix/B′−C loop region, whereas the A481 mutation triggered the conformation changes of its adjacent residues including Phe209 and Phe480 via an indirect manner to affect the substrate binding. Neither mutation significantly changed the substrate binding orientations but did distort the shape of the Phe-cluster in the active site and thus weakened the interactions with coumarin. In addition, very interestingly, the N297S mutation induced a conformation rotation of the side chain of Ser297 during MD simulations. The calculated binding free energies were in good agreement with the relative potency of the experimental binding affinities. To verify our results, which are based on molecule A, molecule B in the crystal structure of 1Z10 was also used to perform the modeling and MD simulations using the same procedure as that of molecule A. A similar protein and substrate dynamic behavior occurred as presented here. Part of the results based on molecule B was provided in the Supporting Information. Our findings will be very useful for understanding the biological function of CYP2A6.

Acknowledgment. We thank Dr. Danni L. Harris for his kindness in offering us the heme parameters. The author (W. Li) thanks the Japanese Society for the Promotion of Science (JSPS) research fellowship for foreign researchers. This work was supported by the Grant-in-Aid for Scientific Research from JSPS. The author (Y. Tang) also gratefully acknowledges financial support from National Natural Science Foundation of China (Grants 20572023).

Supporting Information Available: Part of the results based on molecule B of the crystal structure of the WT-coumarin complex was presented in the supporting material. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Rendic, S. Summary of information on human CYP enzymes: Human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83–448.

(2) Guengerich, F. P. Human Cytochrome P450 enzymes. In *Cytochrome P450:Strucuture, Mechanism, and Biochemistry*, 3rd ed.; Ortiz de Montellano, R. R., Ed.; Klumer Academic/Plenum Publishers: New York, 2005; pp 377−531.

(3) Yun, C. H.; Shimada, T.; Guengerich, F. P. Purification and characterization of human liver microsomal cytochrome P-450 2A6. *Mol. Pharmacol.* **1991**, *40*, 679–685.

(4) Pelkonen, O.; Rautio, A.; Raunio, H.; Pasanen, M. CYP2A6: a human coumarin 7-hydroxylase. *Toxicology* **2000**, *144*, 139–147.

(5) Cashman, J. R.; Park, S. B.; Yang, Z. C.; Wrighton, S. A.; Jacob, P., 3rd; Benowitz, N. L. Metabolism of nicotine by human liver microsomes: Stereoselective formation of *trans*-nicotine *N*′-oxide. *Chem. Res. Toxicol.* **1992**, *5*, 639–646.

(6) Idle, J. R. CYP2A6 polymorphism, nicotine, and environmental nitrosamines. *Lancet* **1999**, *353*, 2073.

(7) Yamazaki, H.; Inui, Y.; Yun, C. H.; Guengerich, F. P.; Shimada, T. Cytochrome P450 2E1 and 2A6 enzymes as major catalysts for metabolic activation of *N*-nitrosodialkylamines and tobacco-related nitrosamines in human liver microsomes. *Carcinogenesis* **1992**, *13*, 1789–1794.

(8) Oscarson, M.; Gullsten, H.; Rautio, A.; Bernal, M. L.; Sinues, B.; Dahl, M. L.; Stengard, J. H.; Pelkonen, O.; Raunio, H.; Ingelman-Sundberg, M. Genotyping of human cytochrome P450 2A6 (CYP2A6), a nicotine C-oxidase. *FEBS Lett.* **1998**, *438*, 201–205.

(9) Yano, J. K.; Denton, T. T.; Cerny, M. A.; Zhang, X. D.; Johnson, E. F.; Cashman, J. R. Synthetic inhibitors of cytochrome P-450 2A6: Inhibitory activity, difference spectra, mechanism of inhibition, and protein cocrystallization. *J. Med. Chem.* **2006**, *49*, 6987–7001.

(10) Ikeda, K.; Yoshisue, K.; Matsushima, E.; Nagayama, S.; Kobayashi, K.; Tyson, C. A.; Chiba, K.; Kawaguchi, Y. Bioactivation of tegafur to 5-fluorouracil is catalyzed by cytochrome P-450 2A6 in human liver microsomes in vitro. *Clin. Cancer Res.* **2000**, *6*, 4409–4415.

(11) Torchin, C. D.; McNeilly, P. J.; Kapetanovic, I. M.; Strong, J. M.; Kupferberg, H. J. Stereoselective metabolism of a new anticonvulsant drug candidate, losigamone, by human liver microsomes. *Drug Metab. Dispos.* **1996**, *24*, 1002–1008.

(12) Rautio, A. Polymorphic CYP2A6 and its clinical and toxicological significance. *Pharmacogen. J* **2003**, *3*, 5–7.

(13) Gillam, E. M.; Aguinaldo, A. M.; Notley, L. M.; Kim, D.; Mundkowski, R. G.; Volkov, A. A.; Arnold, F. H.; Soucek, P.; DeVoss, J. J.; Guengerich, F. P. Formation of indigo by recombinant mammalian cytochrome P450. *Biochem. Biophys. Res. Commun.* **1999**, *265*, 469–472.

(14) Nakamura, K.; Martin, M. V.; Guengerich, F. P. Random mutagenesis of human cytochrome p450 2A6 and screening with indole oxidation products. *Arch. Biochem. Biophys.* **2001**, *395*, 25–31.

(15) Guengerich, F. P.; Sorrells, J. L.; Schmitt, S.; Krauser, J. A.; Aryal, P.; Meijer, L. Generation of new protein kinase inhibitors utilizing cytochrome p450 mutant enzymes for indigoid synthesis. *J. Med. Chem.* **2004**, *47*, 3236–3241.

(16) Kim, D.; Wu, Z. L.; Guengerich, F. P. Analysis of coumarin 7-hydroxylation activity of cytochrome P450 2A6 using random mutagenesis. *J. Biol. Chem.* **2005**, *280*, 40319–40327.

(17) Yano, J. K.; Hsu, M. H.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen. *Nat. Struct. Mol. Biol.* **2005**, *12*, 822–823.

(18) Sansen, S.; Hsu, M. H.; Stout, C. D.; Johnson, E. F. Structural insight into the altered substrate specificity of human cytochrome P450 2A6 mutants. *Arch. Biochem. Biophys.* **2007**, *464*, 197–206.

(19) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(20) Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R. D. Prediction of binding modes for ligands in the cytochromes P450 and other heme-containing proteins. *Proteins* **2005**, *58*, 836–844.

(21) Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking. *J. Med. Chem.* **2008**, *51*, 7469–7477.

(22) Guengerich, F. P. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.* **2001**, *14*, 611–650.

(23) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(24) 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* **1988**, *37*, 785–789.

(25) Gilson, M. K.; Sharp, K. A.; Honig, B. H. Calculating the electrostatic potential of molecules in solution. *J. Comput. Chem.* **1988**, *9*, 327–335.

(26) Harris, D. L.; Park, J. Y.; Gruenke, L.; Waskell, L. Theoretical study of the ligand-CYP2B4 complexes: Effect of structure on binding free energies and heme spin state. *Proteins* **2004**, *55*, 895–914.

(27) Frisch, M. J. T., 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, 2004.

(28) Hui, Li; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and interpretation of protein pK_a values. *Proteins* **2005**, *61*, 704–721.

(29) Case, D. A.; Cheatham, T. E., 3rd; 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.

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

(31) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(32) 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.

(33) Rychaert, 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.

(34) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities:measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884–1897.

(35) Rydberg, P.; Rod, T. H.; Olsen, L.; Ryde, U. Dynamics of water molecules in the active-site cavity of human cytochromes P450. *J. Phys. Chem. B* **2007**, *111*, 5445–5457.

(36) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discov. Des.* **2000**, *18*, 113–135.

(37) Wang, J. M.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, *123*, 5221–5230.

(38) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules:

Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(39) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217–230.

(40) 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.

(41) Brooks, B. R.; Janezic, D.; Karplus, M. Harmonic analyis of large systems. *J. Comput. Chem.* **1995**, *16*, 1522–1553.

(42) Weis, A.; Katebzadeh, K.; Soderhjelm, P.; Nilsson, I.; Ryde, U. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *J. Med. Chem.* **2006**, *49*, 6596–6606.

(43) Feenstra, K. A.; Starikov, E. B.; Urlacher, V. B.; Commandeur, J. N.; Vermeulen, N. P. Combining substrate dynamics, binding statistics, and energy barriers to rationalize regioselective hydroxylation of octane and lauric acid by CYP102A1 and mutants. *Protein Sci.* **2007**, *16*, 420–431.

(44) de Graaf, C.; Oostenbrink, C.; Keizers, P. H.; van Vugt-Lussenburg, B. M.; Commandeur, J. N.; Vermeulen, N. P. Free energies of binding of *R*- and *S*-propranolol to wild-type and F483A mutant cytochrome P450 2D6 from molecular dynamics simulations. *Eur. Biophys. J.* **2007**, *36*, 589–599.

(45) de Graaf, C.; Oostenbrink, C.; Keizers, P. H.; van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2006**, *49*, 2417–2430.

(46) Carlsson, J.; Boukharta, L.; Aqvist, J. Combining docking, molecular dynamics and the linear interaction energy method to predict binding modes and affinities for non-nucleoside inhibitors to HIV-1 reverse transcriptase. *J. Med. Chem.* **2008**, *51*, 2648–2656.

(47) Honma, W.; Li, W.; Liu, H.; Scott, E. E.; Halpert, J. R. Functional role of residues in the helix B' region of cytochrome P450 2B1. *Arch. Biochem. Biophys.* **2005**, *435*, 157–165.

(48) Negishi, M.; Uno, T.; Honkakoski, P.; Sueyoshi, T.; Darden, T. A.; Pedersen, L. P. The roles of individual amino acids in altering substrate specificity of the P450 2a4/2a5 enzymes. *Biochimie* **1996**, *78*, 685–694.

(49) Hata, M.; Tanaka, Y.; Kyoda, N.; Osakabe, T.; Yuki, H.; Ishii, I.; Kitada, M.; Neya, S.; Hoshino, T. An epoxidation mechanism of carbamazepine by CYP3A4. *Bioorg. Med. Chem.* **2008**, *16*, 5134–5148.

(50) He, X. Y.; Shen, J.; Hu, W. Y.; Ding, X.; Lu, A. Y.; Hong, J. Y. Identification of Val117 and Arg372 as critical amino acid residues for the activity difference between human CYP2A6 and CYP2A13 in coumarin 7-hydroxylation. *Arch. Biochem. Biophys.* **2004**, *427*, 143–153.

CT900018T

JCTC Journal of Chemical Theory and Computation

QM/MM Simulation on P450 BM3 Enzyme Catalysis Mechanism

Li Tian and Richard A. Friesner*

Department of Chemistry, Columbia University, New York, New York 10027

Received January 21, 2009

Abstract: Using a structure generated by induced fit modeling of the protein−ligand complex, the reaction path for hydrogen atom abstraction in P450 BM3 is studied by means of mixed QM/MM methods to determine the structures and energetics along the reaction path. The IFD structure is suitable for hydrogen atom abstraction at the $\omega-1$ position. The electronic structures obtained are similar to those observed in P450 cam. We show that the barrier for the hydrogen abstraction step from QM/MM modeling is 13.3 kcal/mol in quartet and 15.6 kcal/mol in doublet. Although there is some strain energy present in the ligand, the activation barrier is not dramatically affected. A crystal water molecule, HOH502, plays a role as catalyst and decreases the activation barrier by about 2 kcal/mol and reaction energy by about 3−4 kcal/mol. To achieve reactive chemistry at the remaining experimentally observed positions in the hydrocarbon tail of the ligand, other structures would have to be utilized as a starting point for the reaction. Finally, the present results still leave open the question of whether DFT methods provide an accurate computation of the barrier height in the P450 hydrogen atom abstraction reaction.

1. Introduction

Cytochrome P450 is an essential enzyme for biological functioning in a wide variety of organisms.[1] The primary function of the enzyme is the insertion of an oxygen atom (derived from dioxygen) into a hydrocarbon bond to form an alcohol; other reactions involving oxygen insertion, such as epoxidations, can be catalyzed as well. The insertion of oxygen increases the solubility of lipophilic compounds, and thus makes it easier to excrete them (e.g., through the kidneys in higher organisms). This ability to transform exogenous compounds is of vital importance in mediating the interaction of organisms with their environment.[2−5] Furthermore, interactions with various P450s in the liver often have a critical effect on drug metabolism, drug−drug interactions, and other events of pharmaceutical significance.[6] Consequently, the development of a detailed atomic level understanding of the mechanism of cytochrome P450 functioning is a high priority for computational biochemistry.

The reaction mechanism of P450 has been investigated in a large number of publications,[7−10] primarily focusing on P450 cam, a bacterial enzyme for which extensive structural, spectroscopic, and kinetic data are available.[11,12] Although a great deal of progress has been made in understanding the P450 cam catalytic cycle, there are still puzzling questions concerning the difficulty in trapping the proposed iron oxo intermediate, which is hypothesized to be the species that removes a hydrogen from the substrate and then inserts an oxygen atom.[13,14] Calculated values of the free energy barrier for this reaction are incompatible (too high) with the rapidity of a reaction that would be consistent with the challenges observed in trapping the intermediate. There are various possibilities that can explain this discrepancy, including inaccuracies in the density functional (DFT) electronic structure methods that have invariably been used to model the reaction, and alternative interpretations of the experimental data that do not require a low barrier (including the possibility that some other species is actually the key catalytic form of the enzyme). Investigation of these possibilities is currently ongoing in our laboratory as well as other research groups.

When one moves beyond P450 cam to investigate the reaction mechanisms of other P450s, an additional problem is encountered. In P450 cam, the camphor substrate is perfectly positioned for oxygen insertion relative to the

* Corresponding author e-mail: rich@chem.columbia.edu.

10.1021/ct900040n CCC: $40.75
Published on Web 04/17/2009

putative iron oxo intermediate, with the appropriate hydrogen of the camphor oriented appropriately toward the oxygen of the complex.[11] However, examination of the crystal structures of other P450 species reveals that in many, if not most, cases, the positioning of the ligand in the crystal structure is incompatible with the reported metabolic chemistry.[6,15−19] This observation is actually not terribly surprising. The crystal structures are typically obtained at low temperature, but it is clear that to accommodate a wide range of exogenous species in the active site, induced fit effects, in which the active site residues reorganize, must be quite important.

To investigate this hypothesis, we have carried out several studies of P450 BM3, a bacterial enzyme in which crystallographic studies of ligands have exhibited the discrepancies discussed above. Induced fit modeling[20−22] starting from the PDB structure 1jpz[15] reveals a gating mechanism in which a phenylalanine residue (Phe 87) flips upward, allowing the tail of the fatty acid ligand to move toward the Fe═O moiety.[23] Replica exchange molecular dynamics (REMD) calculations demonstrate that the new conformation (along with a second, closely related conformation generated during the MD simulations) increases in population as the temperature is increased. The results have been confirmed by comparisons with NMR and optical spectroscopic experiments carried out by McDermott and co-workers.[24] These experiments show a temperature-dependent activation of the P450 catalytic cycle, in which movement of Phe87 can be prominently detected. The temperature dependence in the experiments is in good semiquantitative agreement with that predicted from the replica exchange MD simulations.[25,26]

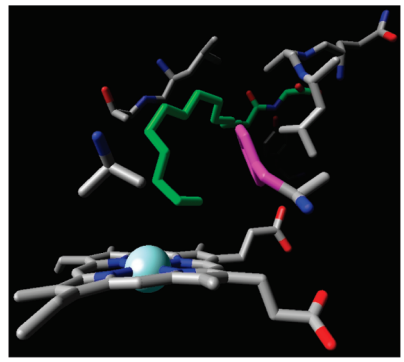
Although these calculations yield structures that can plausibly engage in metabolic chemistry with the catalytic site of P450 BM3, detailed quantum chemical modeling of the P450 BM3 reaction has not yet been performed. In the present paper, we address this question via DFT-based mixed quantum mechanical/molecular mechanics (QM/MM) modeling of the P450 BM3 active site. Starting from the induced fit structure for the complex, we show that the barrier for the rate-determining step of the hydroxylation reaction (hydrogen atom abstraction) is similar in magnitude to that calculated for P450 cam. This result suggests that the problem with calculating the barrier height is consistent for at least two rather different local environments of the catalytic species, suggesting that the explanation of an intrinsic problem with DFT modeling of the reaction may well be at fault. The results also validate the idea of using induced fit modeling, followed by quantum chemical computations, is a useful approach to investigating P450 metabolism in which induced fit effects play a significant role.

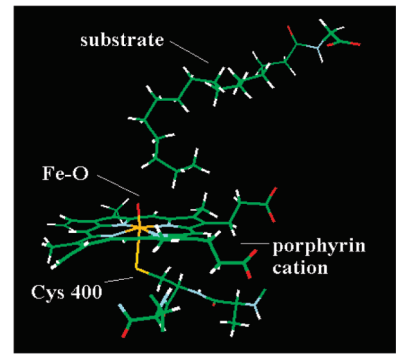
The paper is organized as follows. In section II, we discuss the initial setup of the P450 BM3 induced fit system and briefly outline the computational methods used to both generate the initial structures and carry out the QM/MM calculations. Section III presents both QM and QM/MM results for reactant, transition state, and product structures and for various electronic states of the reacting system and discusses the results from the point of view of both induced fit effects and energetics associated with P450 hydroxylation. Finally, in section IV, the conclusion, the results are summarized and future research directions are indicated.

2. Methods

The structure of the BMP domain of P450 BM3 bound with N-Palmitoylglycine, the substrate, has been determined by X-ray diffraction (1.65 Å, PDB Entry: 1jpz).[15] However, the ω end of substrate in this X-ray structure is too far from the reactive core of the porphyrin oxo species (presumed to be compound I) to make it possible to undergo reactive chemistry. Thus, the reaction cannot be performed starting from this "distal state" and another intermediate with the substrate being bound near the active site, the "proximal state", should exist along the path going from the distal state to the products. An induced fit docking (IFD) approach, a method that can accurately account for both ligand and receptor flexibility, is employed initially to find the possible conformations for the reactive chemistry. IFD, presented by Sherman et al.,[22] combines techniques for docking ligands into a rigid receptor with those for modeling receptor conformational changes, in an iterative fashion. The docking program Glide[20,21,27] is employed to account for ligand flexibility and the refinement module in the Prime program[27] is used to account for receptor flexibility. Tests of the IFD protocol on 21 challenging cross-docking cases taken from the PDB demonstrate that this induced fit method can dock the ligand properly as well as capture the key ligand−protein interactions.[22] In the study by Jovanovic et al.,[23] a docking structure was generated by the induced fit method, as shown in Figure 1a. In this induced fit structure, Phe87 rotates upward so that the ω end of the substrate is allowed to approach the Fe═O core, and the resulting structure has properties consistent with a proximal state. The Prime energy of this structure is only 5.8 kcal/mol higher than that of the nativelike prediction; and the ω end of the substrate is very close to heme iron: $\omega-1$, $\omega-2$, and $\omega-3$ carbons are at distances ranging from 3.6 to 5.8 Å from the iron. Thus, this structure, or structures very close to it, may be conjectured as reactive conformations.

Replica exchange molecular dynamics (REMD) simulations were also employed to study this system in order to explore the conformational equilibrium of P450 BM3 system. Replica exchange (or parallel tempering) molecular dynamics[28,29] involves running a series of constant temperature MD simulations at different specified temperatures for a certain number of time steps. Each simulation at a specific temperature is one replica system. Attempt to make exchange of conformations between a pair of replicas is made periodically, and the exchange is accepted according to the Metropolis transition probability. This method allows the system to experience jumps between potential basins separated by high barriers more efficiently. In the REMD study by Ravindranathen et al.,[25] the result has shown that proximal state has significant population at a temperature around 300 K, and that the IFD-generated structure is on the pathway of converting from the distal to free ligand proximal state. The free proximal state found in the REMD simulations is slightly different from the IFD-generated conformation; this will be discussed in more detail below.


a.



b.

Figure 1. (a) Binding structure from Glide. The rotation of side chain of Phe87 (shown in magenta) allows the substrate, NPG (show in green), to approach the heme iron for reaction. (b) QM region in the QM/MM simulation, including the heme, the substrate NPG, and Cys400.

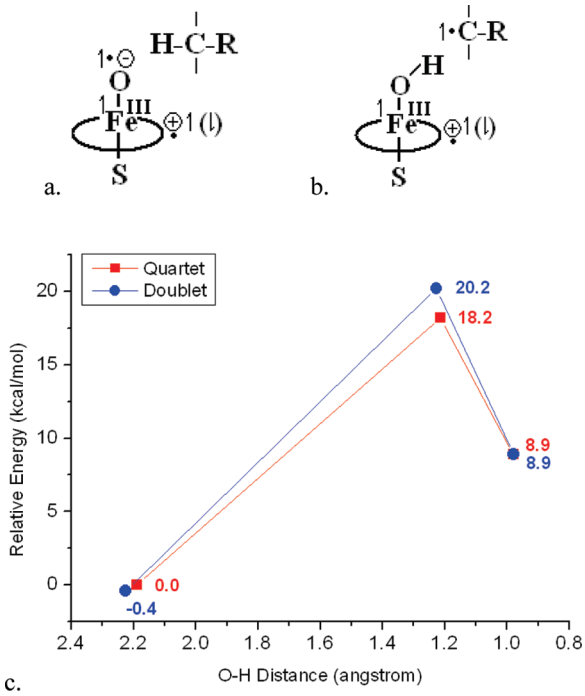


Figure 2. (a) Electronic configurations of reactants in doublet and quartet. (b) Electronic configurations of products in doublet and quartet. (c) Reaction path with relative energy. Energy values are relative to the reactant in quartet. Energy values are in kcal/mol, without ZPE.

On the basis of the reactive induced fit structure, the catalytic mechanism of hydrogen atom abstraction step was investigated in detail in this study. The complicated reaction catalyzed by this transition-metal-containing enzyme involves bond breaking, bond formation, and electron transfer, and there are important effects of the protein/water environment. Thus, a mixed quantum mechanics/molecular mechanics technology is mandatory to capture the detail of the chemistry at the active site and its interaction with the environment. In this study, the structure of our reactant, compound I, was built on the basis of the reactive docking structure from the induced fit method and then modeled with QM/MM as implemented in QSite.[27] Previous QM/MM simulations done in our laboratory, including both enzyme systems and extensive tests on model systems, have demonstrated that the QM/MM protocol in QSite has very small errors associated with the construction of the QM/MM interface.[8,30−36] Using standard protocols for locating reactant, transition state, and product structures and calculating zero point energies, we have also achieved good agreement with experimental free energy barriers (within 2−3 kcal/mol) for a variety of enzymatic reactions via simple transition state theory models.[37] Other effects (longer-range protein motions, quantum tunneling, etc.) are undoubtedly present, but in many if not most cases appear to make relatively minor contributions to experimentally observable energetics. Recent work by Thiel and co-workers has led to similar conclusions.[38]

From the docking structure, the distal oxygen atom was placed above the Fe atom according to the Fe−O distance found in our previous studies of P450 cam. The protein was then soaked in a water shell. As all the crystallographic waters were deleted during Glide docking, we reincorporated the structural waters because some close to the active site may have substantial effect. The water shell was then equilibrated through dynamics simulations using IMPACT.[39] Water molecules beyond 10 Å from the protein were deleted and the oxygen atoms of water molecules 6−10 Å from the protein were frozen throughout the following calculation. The whole system consists of about 22 000 atoms. The structure of the protein was carefully inspected to find titratable residues with unusual protonation states. At the end of this process, a total net charge of −16 e was found. The effect of this net charge on the energetics of the catalysis was tested with a model including counterions but the effect was found to be negligible. This is probably because the charged residues are mainly located at the surface of the protein, near the solvent, and their effect on the active site energetics is screened by solvent (reducing the effective field at long distance by roughly the dielectric constant of water, 80). The reaction was then modeled using QSite. Unrestricted density functional theory was used to model the QM region in conjunction with the B3LYP functional. A polarized basis set (LACVP*) was used for the QM region, as shown in Figure 1b, which includes the heme, the substrate *N*-palmitoylglycine (NPG), and the cysteine (∼160 atoms, ∼1000 basis functions). After geometry optimization was carried out with LACVP*, larger triple-ζ basis sets [cc-pVTZ(-f)] was used for some key atoms in QM region in a single point calculation. Because of the huge size of the QM region (more than 2300 basis functions with cc-pVTZ(-f) basis sets), it was difficult to apply triple-ζ basis sets on the

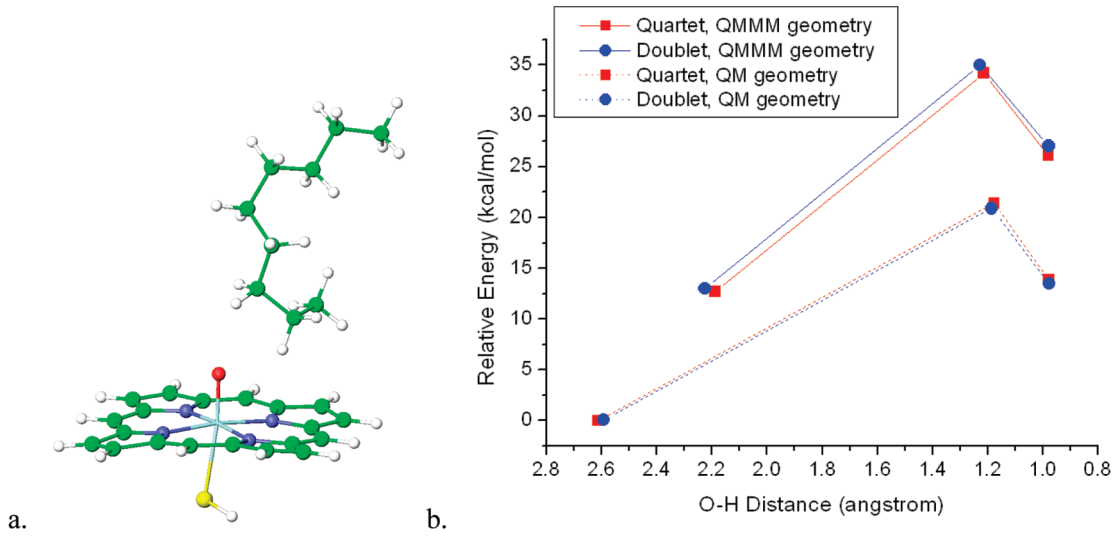


Figure 3. (a) QM model: C_9H_{20} hydrocarbon chain representing the substrate NPG; peripheral substitutes of the heme were all removed and S−H was the model for Cys400. (b) Energy profile.

whole QM region. Thus, we used cc-pVTZ(-f) for the oxygen and four nitrogen atoms bound to Fe as well as the ω terminus of the substrate, used LACV3P* for Fe, and used 6-31G* for others in the test. The energy difference between smaller and bigger basis sets is 0.8 out of 19.2 kcal/mol for activation barrier and 3.3 out of 7.3 kcal/mol for reaction energy. The effect of the larger basis set on the barrier height is minimal, but is somewhat larger for the reaction energy. However, the primary purpose of the present paper is not to quantitatively obtain energetics for the reaction path (particularly given that intrinsic errors in the DFT methods need to be investigated, as will be discussed below), but rather to establish the IFD generated structure as a reasonable model for a ligand pose leading to reactive chemistry. Consequently, to save computer time, we used smaller basis sets, LACVP*, in the following calculations. Finally, the OPLS 1999 force field was used for the MM region, which includes the protein and a shell of water molecules (~22 000 atoms).

Two different QM models were set up, the first to investigate the core active site quantum chemistry in the absence of perturbations, and second to incorporate some effects of protein and solvent environment in a controlled fashion. The most simplified QM model with truncated substrate and heme, as shown in Figure 3a, was used to describe the chemistry of the reaction core. A more extended QM model, the H-capped QM region in QM/MM set up with electrostatic potential from specified point charges, was used to evaluate the effect of protein and water environment. Unrestricted DFT implemented in Jaguar[27] was employed for the two sets of QM model calculations in both doublet and quartet spin states.

3. Results

Mass spectroscopy has shown that the position of hydroxylation is mainly at the $\omega-2$ and $\omega-3$ carbon of the ligand, but also at the $\omega-1$ carbon.[15] Because the $\omega-1$, $\omega-2$, and $\omega-3$ carbons are all reactive positions, the near attack binding structures for each may be different. Therefore, to determine to which reactive position our pose corresponds, we performed a scan along the hydrogen−oxygen distance coordinate for both $\omega-1$ and $\omega-2$ positions. The $\omega-3$ hydrogen was obviously too far for reaction in this binding structure. The $\omega-1$ and $\omega-2$ scan profiles suggest that the docking structure is probably more suitable for hydroxylation at position $\omega-1$, because an activation energy of more than 30 kcal/mol was found in the case of $\omega-2$ hydroxylation. Further sampling of the ligand would be needed to produce structures suitable for investigation of chemistry at the $\omega-2$ and $\omega-3$ positions; we reserve such investigations for another publication.

In the present paper, we report the results of $\omega-1$ hydrogen abstraction and analyze the effect of the protein environment. We first performed QM/MM simulation for reactant and product, and found the transition state, in both quartet and doublet states. The reasonable hydrogen atom abstraction activation barrier obtained in our calculation suggests that the structure obtained from the IFD simulation is realistic. We then discuss QM model calculation results and analyze the effect of the protein and water shell, such as strain energy, electrostatic interaction, and intermediate interaction. Because recent publications[40,41] have reported that a crystal water molecule plays an important role as a catalyst in the P450 cam model of hydrogen atom abstraction, the effect of an analogous water molecule was also analyzed to investigate the effect in the P450 BM3 model.

3.1. QM/MM Result in Quartet and Doublet. Relative energies, optimized geometry parameters, and spin density on critical atoms from QM/MM calculations are shown in Table 1. The two transition states were confirmed by additional minimization starting from points very close to the transition state toward the reactant and the product. Frequency calculation also confirmed the two transition states with negative frequencies (-1826 cm^{-1} in quartet and -1784 cm^{-1} in doublet) corresponding to O−H bond vibration mode. Several other negative frequencies with much smaller energy (mainly several tens to hundreds cm^{-1}), corresponding to small conformational fluctuations of the protein, were also found. This may be due to constraints from substrate binding

Table 1. QM/MM Energy, Optimized Geometry Parameters, and Spin Density on Critical Atoms[a]

Quartet, QM/MM						
	G	*E*	O−H	C−H	O−Fe	Fe−S
reactant	0.0	0.0	2.189	1.092	1.618	2.626
transition state	13.3	18.2	1.214	1.343	1.737	2.505
product	7.3	8.9	0.980	2.073	1.790	2.442

	spin					
	Fe	O	C	H	S	Porp.
reactant	1.15	0.85	0.01	0.00	0.30	0.69
transition state	0.93	0.61	0.53	−0.05	0.24	0.71
product	0.88	0.20	0.94	0.01	0.18	0.78

Doublet, QM/MM						
	G	*E*	O−H	C−H	O−Fe	Fe−S
reactant	−0.4	−0.4	2.225	1.092	1.630	2.548
transition state	15.2	20.2	1.228	1.327	1.748	2.480
product	6.7	8.9	0.979	2.080	1.806	2.425

	spin					
	Fe	O	C	H	S	Porp.
reactant	1.24	0.85	0.01	0.00	−0.26	−0.84
transition state	1.06	0.55	0.47	−0.06	−0.21	−0.84
product	1.00	0.11	0.87	0.01	−0.14	−0.85

[a] *G* is the relative free energy after ZPE correction, and *E* is the relative energy without ZPE; both sets of energy values are relative to quartet reactant and in kcal/mol. Distances are in Å and angles are in degrees. Porp. is the porphyrin π-cation radical.

and/or errors in frequencies calculation. The Gibbs free energy values including zero point energy correction are listed in Table 1, too.

Throughout the reaction, Fe is in its ferric state with net spin around 1.0, and the porphyrin ring is always a radical, as shown in configurations a and b in Figure 2. The sulfur in Cys400, as the fifth ligand of Fe, has localized spin density of about 0.2 and actually shares the third unpaired electron with the porphyrin ring. We will discuss this in detail later. During the reaction, the iron oxo moiety abstracts one electron from C−H bond in the substrate, generating a carbon radical. At the same time, the hydrogen atom is transferred from the carbon to the oxo atom. The difference between doublet and quartet states is mainly manifested in the spin of the porphyrin ring; whether or not the porphyrin ring's spin is parallel to the unpaired spin on Fe and oxo determines the multiplicity of the complex.

The reactants of the two spin states have approximately the same energy, with only a 0.4 kcal/mol difference. The free energy of activation barrier for H abstraction is 13.3 kcal/mol in the quartet, whereas that in the doublet is 15.6 kcal/mol, which is 2.3 kcal/mol higher. The reaction energy is 7.3 kcal/mol for the quartet and 7.1 kcal/mol for the doublet. The breaking of the O−H bond contributes similarly, about 5 kcal/mol, in zero point energy (ZPE) correction in both quartet and doublet cases. The energy values without ZPE are shown in Figure 2c. The comparison between the two spin states suggests that the energetics difference between quartet and doublet is noticeable but small, and the reaction paths should both be feasible in the reaction dynamics. This is consistent with simulation results in the P450 cam model.

3.2. QM Model in Vacuum Result. A simplified QM model, as shown in Figure 3a, was investigated in a vacuum using QM/MM-optimized geometries and vacuum-optimized geometries. Because it has been reported that S−H was a better model for cysteine than methyl mercaptan (CH_3−SH),[42] we built the small QM model including S−H, porphyrin with no substituents, and a 9-carbon hydrocarbon chain. The energy profiles of the quartet and doublet states, QM/MM-optimized geometries, and vacuum-optimized geometries are shown in Figure 3b and listed in Table 2.

We assessed the effect of strain imposed by the protein in the small QM model simulation. From QM/MM-optimized geometries, optimization in a vacuum decreases the energy of the system by about 13 kcal/mol. This indicates that the steric strain and other interactions within the binding cavity impose some constraints on the substrate conformation. We also noticed that several negative vibrational frequencies of QM/MM-optimized geometries are eliminated in the vacuum optimization. These vibrational modes include jiggling of the two ends of the substrate chain and wiggling of S−H. The comparison of conformational change during the reaction in both spin states is shown in Figure 4.

The activation barrier values with and without vacuum optimization in the quartet are 21.4 and 21.5 kcal/mol, respectively, and this result suggests that the optimization does not have a significant effect on the activation barrier in quartet, even though the conformational change through the reaction is much more significant in the optimization. In the case of the doublet, the same conformational change was observed in the optimization; however, the activation barrier was decreased from 22.0 to 20.8 kcal/mol, by removing steric strain in the protein pocket. Although constrained on conformation, optimization in a vacuum does not decrease the activation energy very much, which is almost zero in the quartet (21.5 and 21.4 kcal/mol) and 1.2 kcal/mol in the doublet (22.0 and 20.8 kcal/mol). This suggests that the protein pocket does not facilitate the reaction by conformational constraints or imposed strain. The constraints imposed by the protein do not destabilize or stabilize a particular point on the reaction path, but instead limit the conformation to a smaller phase space without changing the relative energetics.

Comparing the QM model results in a vacuum to the QM/MM results within the protein environment, we observe the main differences in the energy profile and spin distributions. The protein decreases the activation barrier by 3.3 kcal/mol in the quartet but only 1.4 kcal/mol in doublet, and the reaction energy is decreased by 4.5 kcal/mol in the quartet and similarly in the doublet by 4.7 kcal/mol. This indicates that the catalysis mechanism in the quartet and doublet share some common characteristics but also have differences, accounting for about 2 kcal/mol. The main difference in spin distribution is how the third unpaired electron is shared between the porphyrin ring and the sulfur atom in Cys400. Because of the stabilization of the sulfur by protein pocket,

Table 2. QM Model in Vacuum: Relative Energy, Optimized Geometry Parameters, and Spin Density on Critical Atoms[a]

Quartet QM/MM						
	E	O−H	C−H	C−O	C−H−O	relative *E*
reactant	0.0	2.189	1.092	3.020	131.0	12.7
transition state	21.5	1.214	1.343	2.502	156.4	34.2
product	13.4	0.980	2.073	2.960	149.6	26.1

	spin					
	Fe	O	C	H	S	Porp.
reactant	1.09	0.92	0.01	0.00	0.52	0.48
transition state	0.89	0.70	0.53	−0.07	0.39	0.56
product	0.83	0.25	0.95	0.00	0.35	0.63

Doublet QM/MM						
	E	O−H	C−H	C−O	C−H−O	relative *E*
reactant	0.0	2.225	1.092	3.055	131.0	13.0
transition state	22.0	1.228	1.327	2.510	158.3	35.0
product	14.0	0.979	2.080	2.963	149.0	27.0

	spin					
	Fe	O	C	H	S	Porp.
reactant	1.18	0.91	0.01	0.00	−0.53	−0.58
transition state	1.04	0.60	0.44	−0.07	−0.36	−0.68
product	0.99	0.10	0.92	0.01	−0.22	−0.79

Quartet QM Model						
	E	O−H	C−H	C−O	C−H−O	relative *E*
reactant	0.0	2.612	1.097	3.534	141.2	0.0
transition state	21.4	1.177	1.362	2.527	168.8	21.4
product	13.9	0.979	2.334	3.251	155.7	13.9

	spin					
	Fe	O	C	H	S	Porp.
reactant	1.08	0.94	0.00	0.00	0.54	0.45
transition state	0.97	0.67	0.57	−0.06	0.46	0.38
product	0.93	0.25	0.98	0.00	0.47	0.40

Doublet QM Model						
	E	O−H	C−H	C−O	C−H−O	relative *E*
reactant	0.0	2.593	1.097	3.527	142.6	0.1
transition state	20.8	1.185	1.328	2.502	169.3	20.9
product	13.4	0.977	2.297	3.205	154.5	13.5

	spin					
	Fe	O	C	H	S	Porp.
reactant	1.20	0.89	0.00	0.00	−0.58	−0.52
transition state	0.99	0.54	0.46	−0.06	−0.42	−0.54
product	0.98	0.11	0.96	0.01	−0.44	−0.60

[a] *E* is energy relative to corresponding reactant in kcal/mol. Relative *E* is energy relative to quartet reactant with QM model optimized geometry.

the net spin on the sulfur is about 0.15−0.3, which is 0.1−0.2 less compared with the net spin in QM model in a vacuum.

4. Discussion

On the basis of the transition state and the energy profile, the proximal binding structure from the IFD calculation is a good conjecture about the starting point (reactant) for hydroxylation on the $\omega-1$ position of the ligand, and both quartet and doublet are feasible reaction paths. In the QM/MM simulation of a similar system P450 cam, Schöneboom et al.[43] reported energy profiles and spin distribution as listed in Table 3. The result is consistent with what we found in P450 BM3, with similar activation barrier and endothermicity. However, the slight but noticeable difference between quartet and doublet energetics is not reported in P450 cam system.

Besides the distal conformational state observed in X-ray structure and the proximal state we used in this paper to predict catalysis, another proximal state was identified in molecular dynamics calculation by Ravindranathan et al.[25] The structure is shown in Figure 5. The ω end of the substrate is very close to the iron atom. The conformational entropy stabilizes this state, and surrounding residues give the substrate the ability to move more freely. However, on the basis of our scanning of hydrogen abstraction potentials, significant rearrangement is needed before hydroxylation can happen on any of $\omega-1$, $\omega-2$, and $\omega-3$ positions.

To evaluate the effect of various contributions to the electrostatic potential of the protein, we performed single-point calculation with QM/MM geometries. The QM model in this series of calculation was the same as the QM region, with H-capped Cys400. Formal charge values and geometry coordinates of all the water molecules and protein residues are extracted from QM/MM-optimized geometries, whereas only the backbone of Ala399 and Ile401, which are connected to QM region, are removed. A single-point QM calculation is then performed with the fully implemented electrostatic potential as in the protein environment. The corresponding energy profile and spin distribution data are shown in Table 4. It is clear from the comparison that modeling the electrostatic potential by point charges can recover almost all the effect of the protein on the wave functions and energetics of species in the catalytic cycle, from decreasing the activation barrier to tuning the spin distribution between the Cys400 sulfur and the porphyrin ring. This indicates that in the case of P450 BM3 hydrogen atom abstraction, the main effect of protein catalysis is to provide an appropriate electrostatic potential. We will discuss several of the most important components of this significant electrostatic potential and the responsible residues. Key residues around the active site are shown pictorially in Figure 6.

We begin with a discussion of the propionate substituents of the heme and their local protein environment. Positively charged residues Lys69 and Arg398 next to the heme peripheral propionates function as a structural anchor for the porphyrin. The Lys and Arg are also important components of the electrostatic environment, lowering the energy levels of the peripheral carboxylate moieties by stabilizing the negative charge on the two propionate groups. When a carboxylate is left as COO− in vacuum without proper screening from either positively charged residue or an equivalent point charge, it cannot maintain the full −1 charge and becomes partly radical with 0.3−0.5 net spin, because compound I is a strong oxidant. Although this destabilization of peripheral substitutes might not show a significant effect on the activation barrier or endothermicity in a single-point

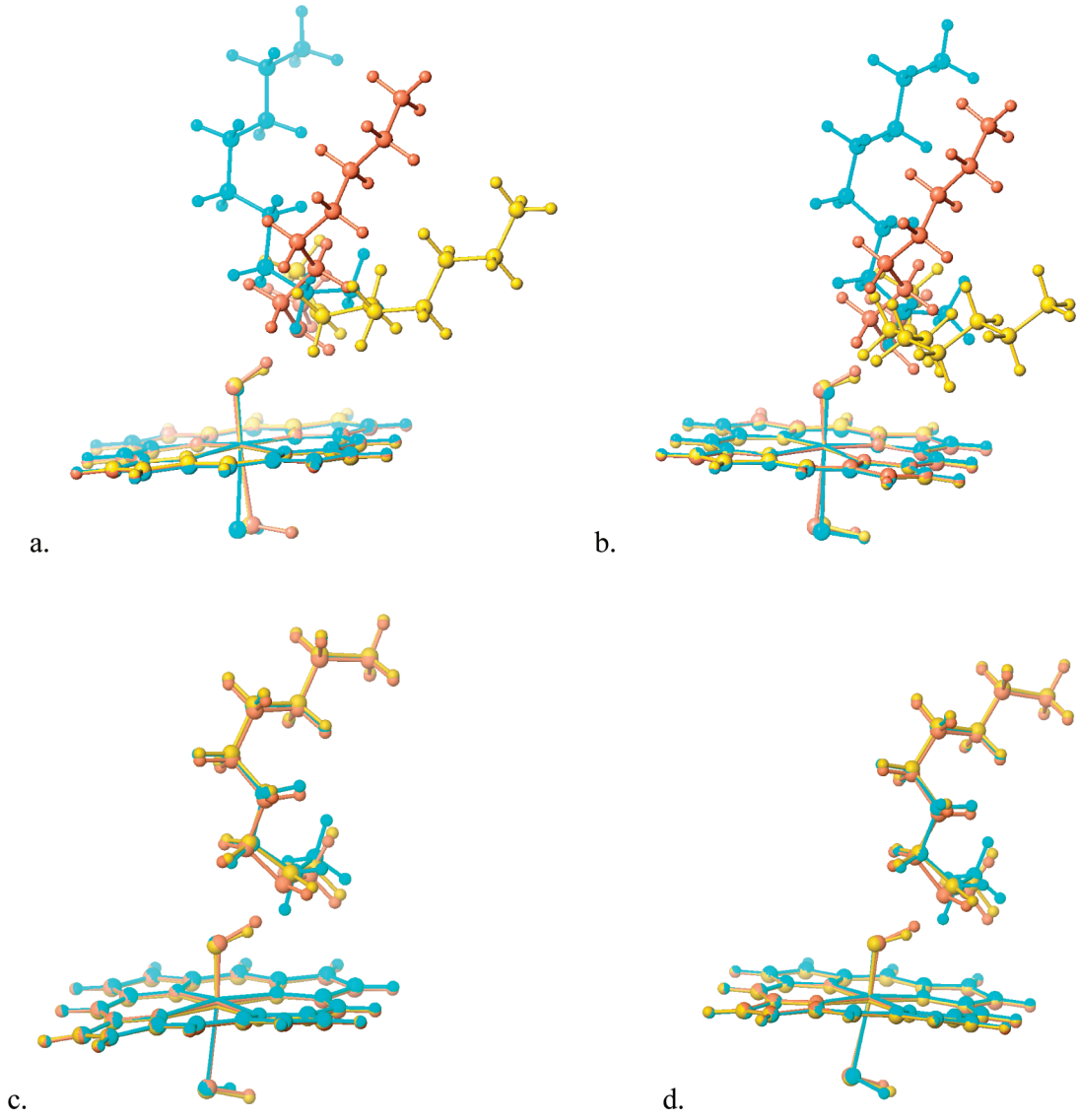


Figure 4. Superposition of reactant, transition state, and product geometries. Reactant is in cyan, transition state is in coral, and product is in goldenrod. Optimized QM model geometries: (a) quartet, (b) doublet. QM/MM geometries: (c) quartet, (d) doublet.

Table 3. Comparisons of Energy Profiles and Spin Distribution in QM/MM Calculation of P450 cam and P450 BM3[a]

		A. E.	R. E.	O−H	spin Fe	spin O	spin C	spin S	spin Porp.
P450 BM3	quartet	18.2	8.9	1.214	0.93	0.61	0.53	0.24	0.71
P450 BM3	doublet	20.6	9.3	1.228	1.06	0.55	0.47	−0.21	−0.84
P450 cam[b]	quartet	21.8	10.5	1.188	0.95	0.61	0.55	0.21	0.72
P450 cam[c]	doublet	21.1	9.4	1.195	1.16	0.55	0.50	−0.26	−0.89
P450 cam[c]	quartet	18.7	6.5	1.225	0.98	0.61	N/A	N/A	0.77

[a] A. E. is the activation barrier in kcal/mol and R. E. is the reaction energy in kcal/mol. Both are before ZPE corrections. Distances and spin density values are of transition state. [b] R2/B2, P450 cam result based on snapshot 40, which did not consider the catalysis effect of crystal water w903; heme peripheral substitute are replaced by H atoms, in QM region.[43] [c] R2/B2, P450 cam result based on snapX incorporating the effect of water903, without heme peripheral substitute in QM region, protonated Asp297.[41]

calculation, it is not natural and not indicated by the experimental protein structure of this model. Thus it is preferable, as a general guideline in QM/MM, to prepare the protein such that the QM part stays globally neutral as to avoid unphysical electron transfer.

However, although mainly because of electrostatic interaction, the H-bond component in the interaction between positive residues and propionates is not negligible. The QM/MM energy difference between two orientations of the amine group in Lys69 is about 8 kcal/mol. The orbital shown in Figure 7a indicates a π orbital of the heme propionate, which forms the π cation H-bond with positively charge amino group. In the lower-energy geometry, the π orbital is almost perpendicular to C═O plane and parallel with N−H σ bond in the amino group of Lys69, whereas in the higher-energy conformation, the angle is about 40°.

There is one crystal water molecule, HOH502, located around the reaction core, with position shown in Figure 6.

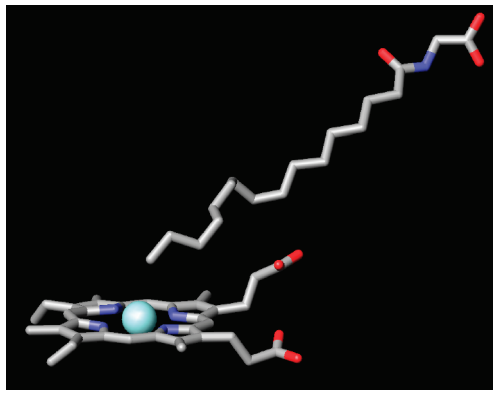


Figure 5. Active site of another proximal state binding conformation predicted by Ravindranathen et al.[25]

Table 4. QM Model with Point Charges: Energy Profiles and Spin Distribution. This Data Can Be Compared with QM/MM Results Listed in Table *1*[a]

Quartet, QM						
	E	spin Fe	spin O	spin C	spin S	spin Porp.
reactant	0.0	1.15	0.85	0.01	0.34	0.65
transition state	19.4	0.94	0.61	0.54	0.29	0.65
product	9.8	0.88	0.20	0.94	0.20	0.77

Doublet, QM						
	E	spin Fe	spin O	spin C	spin S	spin Porp.
reactant	−0.2	1.26	0.85	0.01	−0.31	−0.80
transition state	19.8	1.07	0.55	0.47	−0.25	−0.80
product	9.3	1.02	0.10	0.85	−0.17	−0.80

[a] *E* is the energy relative to quartet reactant, in kcal/mol.

To evaluate the effect of this water molecule, we performed single-point calculation of both QM/MM and QM models in the protein electrostatic environment without HOH502. In the QM/MM calculation, the HOH502 was moved from active site to 60 Å away, which ensured that this water does not impact the active site energetics in a meaningful way. In the QM model calculation, point charges from all of the atoms in the protein and water shell were applied as the electrostatic environment, except for the three points from HOH502. The comparison of the results is shown in Table 5. From these results, it is clear that HOH502 decreases the activation barrier by about 2−3 kcal/mol and the reaction energy by about 3−4 kcal/mol, in both quartet and doublet states, with slightly different quantitative effects depending on calculational settings (this level of uncertainty is, however, negligible compared to other possible errors in the full QM/MM calculations).

The H abstraction reaction, by its nature, increases the number of electrons in heme, the oxidant. The oxo part starts as a negative radical, abstracts one electron and forms an O−H bond; along with the increasing of electrons, the localized spin of the oxo group is decreased and the partial charge on the oxo site becomes more negative. The water HOH502, which is very close to the active site, stabilizes the whole system through its hydrogen bond with oxo, by about 2 kcal/mol in the QM/MM calculations. Moreover, this interaction is stronger with the transition state and product as compared to the reactant. This is because oxo gains more electron density, and becomes more negative, during the course of the reaction, which is favored by this hydrogen-bonded structure. It is clear from data in Table 5 that the electrostatic interaction in the H-bond with HOH502 decreases the spin on oxo by 0.02−0.07 and draws in 0.04−0.08 more negative charge, which is in line with the trend of the reaction. Thus this crystal water functions as a catalyst by stabilizing the transition state and product. From our result in QM model, it is clear that this catalysis effect is mainly an electrostatic interaction: substituting point charges for water molecule can recover the effect.

In the X-ray structure 1JPZ, the B factor of the oxygen atom in HOH502 is 19.98, which is comparable to most atoms in porphyrin ring and significantly smaller than most atoms in the substrate. This indicates that HOH502 is a well-ordered water molecule and its position, as well as interaction with it, should be stable. This observation is also consistent with recent result of calculations on P450 cam. Altun et al.[40,41] reported that in P450 cam, a similar water molecule w903 decreases barrier and endothermicity by about 4 and 6 kcal/mol, respectively, in QM/MM, and 3.0 and 5.2 kcal/mol in QM model calculation. Results of P450 cam system in quartet with similar QM region settings are listed in Table 3. This water provides the most significant catalytic factor of the protein. In the doublet, the electrostatic effect of HOH502 can solely recover the activation barrier difference between the full QM/MM calculation and QM model in a vacuum. In the quartet, besides the catalysis of HOH502, there is still 1.1 kcal/mol stabilization from protein in addition.

In the QM/MM simulation, different electronic configurations with very close energy values under the same geometries were found, indicating that this system has an extremely complicated electronic structure. The main difference between these configurations is how the third single electron is shared between the porphyrin ring and the sulfur atom in Cys400. Our QM/MM results show that lower net spin on sulfur does not always lead to lower energy, and net spin on sulfur in the lowest energy configuration is about 0.2 to 0.3, decreasing along the reaction coordinate. This result is consistent with reported P450 cam values.[43]The spin delocalization, in this case as the electron transfers from the thiolate ligand to the conjugated porphyrin ring, has some effects on the energy profiles, although these are limited from a quantitative point of view. The net spin on the sulfur atom can be as high as 0.5 in the QM model calculation and it can also be as low as about zero in some electronic configurations in the QM/MM calculation, which both lead to higher activation barrier. We estimate that the effect of tuning the net spin on the sulfur atom is less than 1 kcal/mol, on both the activation barrier and the reaction energy.

It is interesting to find out factors stabilizing the sulfur, tuning the spin distribution between the sulfur and the porphyrin, and finally lowering the activation barrier. As we have discussed before, electrostatic potential is the key effect from protein, and thus the question will be which residues or electrostatic factors are responsible. In P450 cam system, the stabilization effect for sulfur is reported to be the

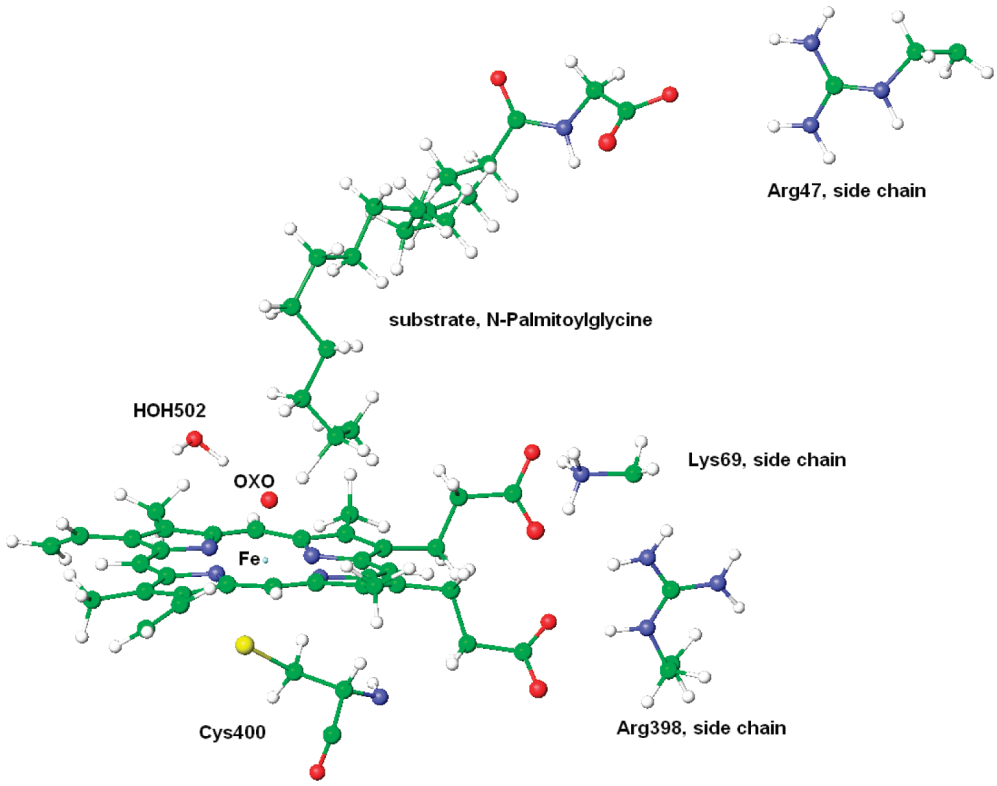


Figure 6. Key residues around the active site of P450 BM3.

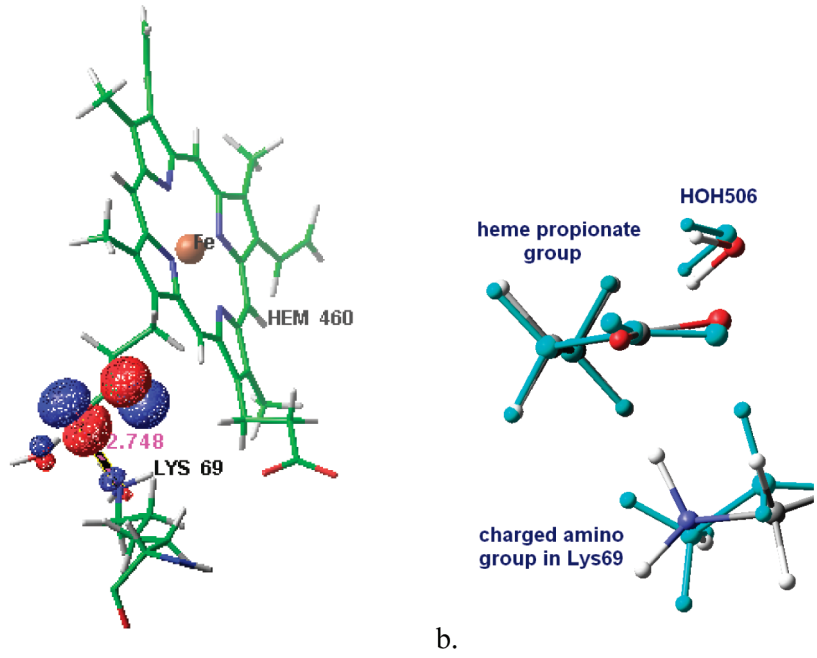


Figure 7. (a) π Orbital of propionate. (b) Superposition of two side chain conformations of Lys69. Higher-energy geometry is in cyan.

H-bonding with neighboring residues Gly359 and Gln360 and the polarization of protein environment, e.g., higher dielectric constant ($\varepsilon = 5.7$).[44,45] In P450 BM3, there are two residues at very similar position showing H-bonding with the sulfur: Gly402 and Gln403. To study whether the same factor plays a role in P450 BM3 system, we investigated the extended QM model as shown in Figure 1b with applied electrostatic potential and dielectric constant. However, neither applying the point charges of the backbone N−H groups of Gly402 and Gln403 nor implementing higher dielectric constant could recover both the spin distribution and the energetics. Moreover, non-negligible energy fluctuations were found in calculations with different model settings. These results indicate that the problem may be more complicated than expected. The propionates connected to conjugated porphyrin ring, as well as the interaction between sulfur atom and backbone of G402Q403 and other surrounding residues may all play a role in this stabilization. The

Table 5. Catalysis Effect of HOH502[a]

Quartet, QM/MM				
	E	spin O	charge O	O−O
reactant	0.0/0.0	0.85/0.87	−0.49/−0.45	2.719
transition state	18.2/20.4	0.61/0.63	−0.67/−0.62	2.667
product	8.9/12.3	0.20/0.22	−0.78/−0.74	2.637

Doublet, QM/MM				
	E	spin O	charge O	O−O
reactant	0.0/0.0	0.85/0.87	−0.49/−0.45	2.715
transition state	20.6/22.6	0.55/0.62	−0.67/−0.59	2.660
product	9.3/13.0	0.11/0.14	−0.79/−0.71	2.632

Quartet, QM			
	E	spin O	charge O
reactant	0.0/0.0	0.85/0.91	−0.46/−0.40
transition state	19.4/22.3	0.61/0.68	−0.65/−0.57
product	9.8/13.9	0.20/0.24	−0.77/−0.69

Doublet, QM			
	E	spin O	charge O
reactant	0.0/0.0	0.85/0.90	−0.47/−0.40
transition state	20.0/23.1	0.55/0.60	−0.64/−0.56
product	9.5/14.0	0.10/0.12	−0.77/−0.69

[a] Values before slash are results with HOH502 and values after slash are results without HOH502. O−O is the distance between oxo and O atom in HOH502, in Å. *E* is energy in kcal/mol relative to reactant of each set. Comparison indicates that HOH502 decreases the activation barrier and endothermicity by about 2−3 and 3−4 kcal/mol, respectively, depending on the different systems.

detail of stabilization of sulfur needs more investigation and will be explored in future research.

Finally, we briefly discuss a recent publication that has studied hydroxylation of a number of alkanes by P450-BM3 using docking, molecular dynamics, and quantum chemical techniques.[46] Although the systems studied in ref 46 and methods used have some similarities to those of the present paper, there are also major differences, e.g., the use of gas phase semiempirical quantum chemistry to compute barrier heights, details of the sampling approaches to generate candidate reactive structures, study of different ligands, and investigation of mutations of the protein. These differences, particularly the very different quantum chemical methods used to estimate barrier heights, make any direct comparisons of the results extremely difficult. As more calculations on flexible P450-based systems are performed, and compared quantitatively to experiment (in ref 46 the comparisons were only qualitative, whereas our own such comparisons presented several difficulties as discussed in the text), it will be possible to achieve a better picture of how well various approaches work in rationalizing P450 chemistry.

5. Conclusion

We have calculated the reaction path for hydrogen atom abstraction in P450 BM3 using a structure generated by induced fit modeling of the protein−ligand complex, and applying mixed QM/MM methods to determine the structures and energetics along the reaction path. The electronic structures obtained are similar to those observed in P450 cam. The IFD structure is suitable for hydrogen atom abstraction at the $\omega-1$ position; although there is some strain energy present in the ligand, it is distributed more or less uniformly in the reactant, transition state, and product, so that the activation barrier is not dramatically affected.

To achieve reactive chemistry at the remaining experimentally observed positions in the hydrocarbon tail of the ligand, other structures would have to be utilized as a starting point for the reaction. The next step in understanding the full range of chemistry in P450 BM3 is to carry out a combination of conformational search and molecular dynamics simulations to produce a suitable ensemble of conformations, and establish that all of the experimentally observed chemistry can be explained via simulations. There are also quantitative questions with regard to the activation free energy required to produce the initial conformations. We have not investigated this issue in the present paper, but a fully quantitative analysis of the thermodynamics of the reactive process would require an assessment of this contribution.

Finally, the present results still leave open the question of whether DFT methods provide an accurate computation of the barrier height in the P450 hydrogen atom abstraction reaction. The fact that the barrier computed here is very similar to that in P450 cam suggests that the explanation for the experimental observation of a very fast reaction is not due to effects of the protein that are somehow being missed, or miscalculated, in the model. Rather, it may be the case that the B3LYP estimation of the barrier height is problematic. The performance of DFT methods for transition-metal-containing systems, and for transition states generally, is far from uniform; systematic errors can be detected in a number of different systems and applications. We have recently been investigating these errors via a novel scheme for introducing empirical localized orbital corrections.[47−49] Application of such methods to the P450 hydrogen atom abstraction is at present not straightforward, but is an area we expect to investigate in future work.

Acknowledgment. This work was supported in part by a grant to R.A.F. from the NIH (GM40526). The authors thank Dr. David Rinaldo for helpful discussions.

References

(1) Ortiz de Montellano, P. R.; De Voss J. J. Substrate Oxidation by Cytochrome P450 Enzymes. In *Cytochrome P450: Structure, Mechanism, and Biochemistry*, 3rd ed.; Ortiz de Montellano, P. R., Ed.; Springer: New York, 2004; pp 183−245.

(2) Capdevila, J. H.; Falck, J. R.; Estabrook, R. W. *Faseb J.* **1992**, *6*, 731–6.

(3) Guengerich, F. P. *J. Biol. Chem.* **1991**, *266*, 10019–22.

(4) Kubota, M.; Sogawa, K.; Kaizu, Y.; Sawaya, T.; Watanabe, J.; Kawajiri, K.; Gotoh, O.; Fujii-Kuriyama, Y. *J. Biochem. (Tokyo)* **1991**, *110*, 232–6.

(5) Porter, T. D.; Coon, M. J. *J. Biol. Chem.* **1991**, *266*, 13469–72.

(6) Williams, P. A.; Cosme, J.; Vinkovic, D. M.; Ward, A.; Angove, H. C.; Day, P. J.; Vonrhein, C.; Tickle, I. J.; Jhoti, H. *Science* **2004**, *305*, 683–686.

(7) Guallar, V.; Baik, M.-H.; Lippard, S. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6998–7002.

(8) Guallar, V.; Friesner, R. A. *J. Am. Chem. Soc.* **2004**, *126*, 8501–8508.

(9) Kamachi, T.; Yoshizawa, K. *J. Am. Chem. Soc.* **2003**, *125*, 4652–4661.

(10) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279–2328.

(11) Schlichting, I.; Berendzen, J.; Chu, K.; Stock, A. M.; Maves, S. A.; Benson, D. E.; Sweet, R. M.; Ringe, D.; Petsko, G. A.; Sligar, S. G. *Science* **2000**, *287*, 1615–1622.

(12) Poulos, T. L.; Finzel, B. C.; Howard, A. J. *J. Mol. Biol.* **1987**, *195*, 687–700.

(13) Davydov, R.; Makris, T. M.; Kofman, V.; Werst, D. E.; Sligar, S. G.; Hoffman, B. M. *J. Am. Chem. Soc.* **2001**, *123*, 1403–1415.

(14) Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. *J. Am. Chem. Soc.* **2007**, *129*, 6204–6209.

(15) Haines, D. C.; Tomchick, D. R.; Machius, M.; Peterson, J. A. *Biochemistry* **2001**, *40*, 13456–13465.

(16) Schoch, G. A.; Yano, J. K.; Wester, M. R.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. *J. Biol. Chem.* **2004**, *279*, 9497–9503.

(17) Williams, P. A.; Cosme, J.; Ward, A.; Angove, H. C.; Vinkovic, D. M.; Jhoti, H. *Nature (London)* **2003**, *424*, 464–468.

(18) Guengerich, F. P. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13565–13566.

(19) Hritz, J.; de Ruiter, A.; Oostenbrink, C. *J. Med. Chem.* **2008**, *51*, 7469–7477.

(20) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(21) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.* **2004**, *47*, 1750–1759.

(22) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.

(23) Jovanovic, T.; Farid, R.; Friesner, R. A.; McDermott, A. E. *J. Am. Chem. Soc.* **2005**, *127*, 13548–13552.

(24) Jovanovic, T.; McDermott, A. E. *J. Am. Chem. Soc.* **2005**, *127*, 13816–13821.

(25) Ravindranathan, K. P.; Gallicchio, E.; Friesner, R. A.; McDermott, A. E.; Levy, R. M. *J. Am. Chem. Soc.* **2006**, *128*, 5786–5791.

(26) Ravindranathan, K. P.; Gallicchio, E.; McDermott, A. E.; Levy, R. M. *J. Am. Chem. Soc.* **2007**, *129*, 474–475.

(27) *Glide, Prime, QSite, Jaguar*; Schrodinger, LLC: Portland, OR.

(28) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 310–321.

(29) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(30) Friesner, R. A. *Drug Discovery Today: Technol.* **2004**, *1*, 253–260.

(31) Friesner, R. A.; Baik, M.-H.; Gherman, B. F.; Guallar, V.; Wirstam, M.; Murphy, R. B.; Lippard, S. J. *Coord. Chem. Rev.* **2003**, *238*−*239*, 267–290.

(32) Gherman, B. F.; Goldberg, S. D.; Cornish, V. W.; Friesner, R. A. *J. Am. Chem. Soc.* **2004**, *126*, 7652–7664.

(33) Gherman, B. F.; Lippard, S. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2005**, *127*, 1025–1037.

(34) Guallar, V.; Jacobson, M.; McDermott, A.; Friesner, R. A. *J. Mol. Biol.* **2004**, *337*, 227–239.

(35) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *Chem. Phys. Lett.* **2000**, *321*, 113–120.

(36) Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **1999**, *20*, 1468–1494.

(37) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

(38) Thiel, W. Presented at the *Theoretical Biochemistry: Methods and Applications Symposium*, Stockholm, May 14−17, 2008.

(39) Banks, J. L.; Beard, H. S.; Cao, Y.; 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. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(40) Altun, A.; Guallar, V.; Friesner, R. A.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2006**, *128*, 3924–3925.

(41) Altun, A.; Shaik, S.; Thiel, W. *J. Comput. Chem.* **2006**, *27*, 1324–1337.

(42) Ogliaro, F.; Cohen, S.; Filatov, M.; Harris, N.; Shaik, S. *Angew. Chem., Int. Ed.* **2000**, *39*, 3851–3855.

(43) Schoneboom Jan, C.; Cohen, S.; Lin, H.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2004**, *126*, 4017–34.

(44) Ogliaro, F.; Cohen, S.; de Visser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2000**, *122*, 12892–12893.

(45) de Visser, S. P.; Ogliaro, F.; Sharma, P. K.; Shaik, S. *J. Am. Chem. Soc.* **2002**, *124*, 11809–11826.

(46) Feenstra, K. A.; Starikov, E. B.; Urlacher, V. B.; Commandeur, J. N. M.; Vermeulen, N. P. E. *Protein Sci.* **2007**, *16*, 420–431.

(47) Friesner, R. A.; Knoll, E. H.; Cao, Y. *J. Chem. Phys.* **2006**, *125*, 124107/1–124107/24.

(48) Knoll, E. H.; Friesner, R. A. *J. Phys. Chem. B* **2006**, *110*, 18787–18802.

(49) Rinaldo, D.; Tian, L.; Harvey, J. N.; Friesner, R. A. *J. Chem. Phys.* **2008**, *129*, 164108/1–164108/23.

CT900040N

JCTC Journal of Chemical Theory and Computation

Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? 2. Crystal Structure Prediction

Panagiotis G. Karamertzanis,*,†,‡ Andrei V. Kazantsev,† Nizar Issa,‡ Gareth W.A. Welch,‡ Claire S. Adjiman,† Constantinos C. Pantelides,† and Sarah L. Price‡

Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, SW7 2AZ, United Kingdom, and Department of Chemistry, University College London, 20 Gordon Street, London, WC1H 0AJ, United Kingdom

Received October 13, 2008

Abstract: We report a multistage lattice energy minimization methodology for generating stable packing arrangements of cocrystals containing flexible molecules. In the first approximation, the intermolecular electrostatic interactions are modeled with atomic charges and the molecular deformation energy is interpolated over a set of precomputed quantum mechanical values. At subsequent stages, the accuracy is improved by first using analytically rotated and then conformation-dependent multipole moments, computed from the isolated-molecule charge density, and "on-the-fly" quantum mechanical calculations to compute the intramolecular deformation energy. This multistage approach increases the efficiency of the search and establishes the molecule-dependent error due to the atomic charge representation of the charge density and the neglect of the conformational dependence of atomic multipole moments. The methodology is used to study the lattice energy landscapes of the cocrystals of 4-aminobenzoic acid with 2,2′-bipyridine and 4-nitrophenylacetic acid, as well as the single-component crystals. All single-component, experimentally determined crystal structures within the scope of the search were found at, or very close to, the global minimum. The experimental cocrystal with 2,2′-bipyridine is also predicted to be among the most stable packing arrangements. On the contrary, the lattice energy landscape of the cocrystal with 4-nitrophenylacetic acid contains several low energy structures that are more stable than the experimentally observed form and have different hydrogen bonding motifs. Overall, the methodology can provide worthwhile crystal energy landscapes for multicomponent organic solids and thereby contribute to understanding cocrystal formation.

Introduction

The design of multicomponent crystals with desired properties,[1] and hence with specific molecular arrangements in the solid state, can be seen as the supramolecular analogue of chemical synthesis. However, while chemical synthesis has set the relationship between reactivity and chemical structure on a firm theoretical footing, supramolecular synthesis turns out to be a much more elusive task despite recent progress.[2,3] Although cocrystal formation can often be anticipated on the basis of the complementarity in hydrogen bonding capabilities of the component molecules,[4] our recent analysis of the cocrystals of three pharmaceutically acceptable coformers showed that a significant proportion of the cocrystals did not have the anticipated hydrogen bonding motif.[5] Cocrystals are expected to be formed if there exist packing arrangements containing both molecules that are thermody-

* Corresponding author. E-mail: p.karamertzanis@imperial.ac.uk.
† Imperial College London.
‡ University College London.

10.1021/ct8004326 CCC: $40.75
Published on Web 04/08/2009

namically more stable than the pure component crystals. Hence, the development of algorithms[6] to predict the structure and thermodynamic stability of single and multicomponent crystals can assist the screening for cocrystals, by showing which, if any, arrangements of the two molecules are thermodynamically plausible. Such algorithms can also promote our understanding of the formation of molecular salts,[7,8] solvates,[9] and hydrates[10,11] and can also be used to rationalize crystallization with more than one chemically identical molecule in the asymmetric unit ($Z' \geq 2$).[12−18]

The most widely used approach for predicting organic crystal structures is to minimize the lattice energy of a large number of systematically or randomly generated candidate structures.[6,19−22] In this approach, the experimentally observed structure is assumed to correspond to the most stable packing arrangement. The identification of polymorphs as low-lying minima in a computational energy landscape is more likely to match physical reality when the entropic contributions to the free energy are also included.[23,24] Such calculations are usually successful in locating the experimentally determined polymorphs, although the rank of the generated structures depends crucially on the model for the crystal energy. Successful predictions[6,25] have been obtained by refining crude search structures with a more accurate energy model, based on either dispersion-corrected periodic density functional methods[26−28] or accurate models to describe the intermolecular contributions to the lattice energy, U_{inter}, and the energy cost of the molecular distortions, ΔE_{intra}, in the crystal.[29−32] It has been established that in the case of hydrogen bonded systems, the predicted relative stability of the structures is very sensitive to the model for the electrostatic contribution to U_{inter}, with the best results obtained using distributed multipole models.[33] The predicted lattice energies often depend significantly on the fine details of the molecular conformation, such as amino group rotation and pyramidalization,[34] which should ideally be optimized simultaneously with the lattice geometry.[29,31] For larger conformational changes, it is also necessary to consider the effect of the molecular conformation on the molecular charge density and consequently the model for the intermolecular interactions.[35−37]

This paper addresses the challenge of using such lattice energy evaluations to predict the range of possible crystal structures for a cocrystal[38] of assumed stoichiometry and for the component molecules and their relative lattice energies, without relying on experimental information. An appealing attribute of the method is that some conformational flexibility, such as rotation around a selected set of single bonds, is modeled from the outset, thus reducing the need to perform multiple searches with rigid conformers.[39,40] The approach is not limited to neutral molecules and is independent of whether the components are solids or liquids at ambient conditions. Therefore, it can be applied to the crystal structure prediction of salts, solvates, and other systems with more than one chemically identical molecule in the asymmetric unit. Finally, unlike an approach previously used for diastereomeric salts[8] and a monohydrate,[10] the method makes no assumptions about the hydrogen bonding between the crystallographically independent entities. This is likely to be particularly important for cocrystals, as it could lead to the identification of unexpected packing motifs that are thermodynamically competitive with the anticipated hydrogen bonding motifs used in cocrystal design.[2,3]

The computational requirements are kept manageable by using three minimization stages of increasing computational cost and accuracy. The first step uses CrystalPredictor[41,42] to generate a large number of crystal structures and to provide a first estimate of their relative stability. It uses an isotropic intermolecular potential for U_{inter} and estimates of ΔE_{intra} interpolated over a set of precomputed ab initio conformational energies. The computational efficiency and parallel implementation of this algorithm make it suitable for searching complex lattice energy surfaces such as those found for cocrystals, salts, and hydrates.

In the second step, a fast reminimization of the most stable unique structures produced by CrystalPredictor is carried out by replacing the point charges with a distributed multipole model to improve the quality of the representation of the electrostatic interactions, particularly the directionality of hydrogen bonds and $\pi-\pi$ stacking. A new algorithm, referred to as DMAflex-Quick, is used for this. It is computationally efficient due to the analytical rotation of the atomic multipole moments carried out to approximate the variation in molecular charge density with conformation. As at stage 1, precomputed interpolation tables are used for the computation of ΔE_{intra}.

At the final stage, the modeling accuracy is improved further by directly computing, with quantum mechanical calculations, the intramolecular energy, and molecular charge density for each conformation that arises in the lattice energy minimization. DMAflex[29] is used for this. The development of the intermediate DMAflex-Quick stage was prompted by the observation that the energy reranking by improving the representation quality of the molecular charge density is often so large that a complete CrystalPredictor search for cocrystals would require an impractically large number of quantum mechanical calculations for the DMAflex minimization of a sufficient number of crystal structures.

To test and illustrate the new algorithm, we apply it to the cocrystals formed by 4-aminobenzoic acid with 2,2′-bipyridine or 4-nitrophenylacetic acid and all three independent single-component crystals. These systems were selected from the set of 26 cocrystals used in part 1 of this paper[5] to provide a stringent test for the algorithm. The total number of degrees of freedom used in the search exceeds 20; this was previously found to be the upper limit, beyond which random UPACK searches frequently missed low energy minima.[11]

There are two known 4-aminobenzoic acid polymorphs. The high temperature polymorph α[43] ($Z' = 2$) comprises $R_2^2(8)$ dimers formed by inversion-related molecules and is kinetically favored at low temperatures under most crystallization conditions (see part 1). The hydrogen bond motif in polymorph β[44,45] ($Z' = 1$) is totally different and based on four-membered rings with alternating carboxylic and amino groups. The 4-aminobenzoic acid:4-nitrophenylacetic acid cocrystal has heterodimers adopting the same $R_2^2(8)$ carboxylic acid hydrogen bonding motif as the α polymorph.

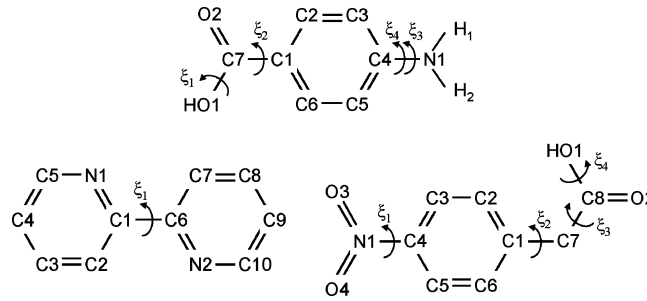
However, the cocrystals of 4-amino- and 4-nitrobenzoic acid with 2,2′- and 4,4′-bipyridine all contain carboxyl$\cdots N_{pyridine}$ hydrogen bonds.[46] In all four of these cocrystals, bipyridine adopts a highly symmetric planar conformation and lies at either an inversion center or a 2-fold axis leading to a 1:2 stoichiometric ratio.

The first question we address is whether the searches are sufficiently complete to consistently find all experimental structures. Then, we analyze how the improvements in the model for the lattice energy change the relative stability of the crystal structures from one stage to the next. This allows us to assess the reranking due to the earlier approximations in the molecular structure, conformational energy, and intermolecular electrostatic interactions. The lattice energy landscapes of the component molecules are examined for the presence of distinct, yet equally thermodynamically favorable, packings. Such crystal energy landscapes, showing that the molecule has many different ways of packing, have been linked[47] to polymorphism and promiscuity in solvate and cocrystal formation.[48] Similarly, by examining the plurality of hydrogen bonding motifs in the cocrystal searches, we discuss whether the experimental structure has any definite energetic advantage over alternatives, and the degree to which cocrystals can be polymorphic. The stability of the low-energy predicted cocrystal structures is finally contrasted with the experimental and theoretically predicted crystal structures of the component molecules to determine whether cocrystal formation could have been predicted a priori.

Computational Methodology. *Extent of Molecular Flexibility.* Forces between molecules are generally weak compared with the forces between covalently bonded atoms. This explains the scarcity of high-energy conformers in experimentally determined crystal structures,[49−51] (There are exceptions for molecules where an intramolecular hydrogen bond in the gas-phase changes to an intermolecular hydrogen bond in the crystal structures[52,53]). Hence, molecular flexibility can often be limited to the distortion of a small number of "flexible" torsion angles denoted as $\boldsymbol{\xi}^{f}$ in the regions around the local conformational-energy minima. If the barrier to interconvert one conformational minimum to another is large, compared with the small energy differences of likely polymorphs, it is both possible and convenient to treat the conformational regions independently.[42]

In Scheme 1, we show the flexible degrees of freedom $\boldsymbol{\xi}^{f}$ that were varied during both the generation of crystal structures and the lattice energy minimizations. In Table 1, we compare the values of $\boldsymbol{\xi}^{f}$ in the experimental crystal structures and in the HF/6−31G(d,p) global conformational minima to assess the effect of the intermolecular forces on the conformations. On the basis of a series of one-dimensional gas-phase quantum mechanical relaxed scans, we modeled conformational changes in the region for which intramolecular energies do not exceed 20−30 kJ mol^{-1} more than the gas-phase global conformational minimum (shown as "range considered in CSP" in Table 1). We confirmed that the resulting region for 4-aminobenzoic acid includes all its conformations found in its 45 crystal structures in the Cambridge Structural Database[54] (CSD hereafter). The CSD analysis showed that the carboxylic group can be displaced by up to 15° ($-15° < \xi_2 < 15°$) from the benzene plane, while the variation in the amine group pyramidalization is greater. We ascertained that the crystal structure prediction searches did not generate any energetically plausible crystal structures containing molecular conformations that lie too close to the border of the conformational region modeled.

Scheme 1. Molecular Structures of 4-Aminobenzoic Acid (top), 2,2′-Bipyridine (bottom left), and 4-Nitrophenylacetic Acid (bottom right)[a]



[a] The flexible torsion angles varied in the generation of crystal structures and lattice energy minimizations are labeled (4-aminobenzoic acid: $\xi_1 \equiv$ H−O1−C7−C1, $\xi_2 \equiv$ O1−C7−C1−C6, $\xi_3 \equiv$ H2−N1−C4−C5, $\xi_4 \equiv$ H1−N1−C4−H2; 2,2′-bipyridine: $\xi_1 \equiv$ N1−C1−C6−N2; and 4-nitrophenylacetic acid: $\xi_1 \equiv$ O3−N1−C4−C3, $\xi_2 \equiv$ C8−C7−C1−C2, $\xi_3 \equiv$ O1−C8−C7−C1, $\xi_4 \equiv$ H−O1−C8−O2). Torsion ξ_4 of 4-aminobenzoic acid is an improper dihedral determining the amine group pyramidalization.

There is a subtle point about molecular symmetry that needs to be considered for cocrystals. The conformational region for 4-aminobenzoic acid contains only positive amine pyramidalization torsion angles (torsion ξ_4) and only describes one of the enantiomers. In the case of single-component crystals, this allows us to reduce the cost of crystal structure prediction without any loss of generality because the other enantiomer will be automatically generated by inversion or mirror plane symmetry operations in racemic space groups, and either enantiomer can be used in chiral space groups. However, the use of only one of the enantiomers of each coformer molecule results in an incomplete search for cocrystals in chiral space groups. Hence, for the cocrystal searches both enantiomers were represented in the conformational regions of both cocrystal formers; thus, the symmetry was exploited only for the purposes of constructing the grid of intramolecular energy values and gradients used by CrystalPredictor and DMAflex-Quick (see section 3.2). Restraining the 4-nitrophenylacetic acid torsion angle ξ_2 to positive values does not limit the scope of the cocrystal search because the molecule is symmetric under half-circle rotation of the nitrobenzene ring with respect to the propionic acid moiety.

Generation of Hypothetical Crystal Structures (CrystalPredictor). Hypothetical crystal structures were generated in 13 densely populated space groups $P1$, $P\bar{1}$, $P2_1$, $P2_1/c$, $P2_12_12$, $P2_12_12_1$, $Pna2_1$, $Pca2_1$, $Pbca$, $Pbcn$, $C2/c$, Cc, and $C2$ using low discrepancy (Sobol') sequences.[55] This was achieved by varying the lattice variables **X**, which comprise the cell lengths, the cell angles, the position and orientation of one of the fragments (base fragment) for each

Table 1. Flexible Torsion Angle Values in the Experimental Crystal Structures and the Gas-Phase HF/6-31G(d,p) Global Minimum for 4-Aminobenzoic Acid, 2,2′-Bipyridine, and 4-Nitrophenylacetic Acid[a]

	ξ_1	ξ_2	ξ_3	ξ_4
	4-Aminobenzoic Acid			
polymorph α (AMBNAC01[43]), $Z' = 2$	−11.8°	−3.4°	+24.3°	+144.6°
	−8.1°	+1.9°	+27.8°	+142.9°
polymorph β (AMBNAC04[45])	−3.8°	+9.3°	+26.8°	+124.6°
cocrystal with 2,2′-bipyridine (DAQYUQ[46])	+4.8°	−0.7°	+20.8°	+141.6°
cocrystal with 4-nitrophenylacetic acid (RILJOL[97])	+7.7°	+4.5°	+13.9°	+164.5°
HF/6-31G(d,p) gas phase	−0.1°	+0.2°	+22.4°	+137.5°
range considered in CSP	−40° to +40°	−30° to +30°	−50° to +50°	+90° to +180°
number of conformations calculated	5	5	7	7
	2,2′-Bipyridine			
pure-component crystal (BIPYRL04[98])	+180.0°			
cocrystal with 4-aminobenzoic acid (DAQYUQ[46])	+180.0°			
HF/6-31G(d,p) gas phase	+180.0°			
range considered in CSP	+0° to +360°			
number of conformations calculated	36			
	4-Nitrophenylacetic Acid			
pure-component crystal (SEMTAF01[99])	−10.4°	+72.3°	+169.2°	−3.1°
cocrystal with 4-aminobenzoic acid (RILJOL[97])	−4.3°	+43.6°	−154.8°	−2.4°
HF/6-31G(d,p) gas phase	+0.4°	+65.7°	+166.6°	−0.3°
range considered in CSP	−50° to +50°	+20° to +160°	−180° to +180°	−40° to +40°
number of conformations calculated	3	9	17	3

[a] The range of torsion angle values considered in the cocrystal structure prediction and the number of points used to construct the intramolecular energy Hermite interpolants are also given.

molecule in the asymmetric unit, and the flexible torsions $\boldsymbol{\xi}^f$, the latter being restricted to the ranges shown in Table 1. The rest of the degrees of freedom, hereafter referred to as "rigid" and denoted $\boldsymbol{\xi}^r$, were frozen at their values in the HF/6−31G(d,p) global conformational minimum, i.e., the molecules were treated as a set of rigid fragments linked with single bonds around which rotation is permitted.

The generation of hypothetical structures was carried out using CrystalPredictor.[41,42] The asymmetric unit of the 4-aminobenzoic acid:2,2′-bipyridine cocrystal contains one 4-aminobenzoic acid molecule and half-a 2,2′-bipyridine molecule and cannot be located in a search with 1:1 stoichiometry. Hence, we also performed a limited search using one 2,2′-bipyridine molecule and both enantiomers of 4-aminobenzoic acid in the asymmetric unit in space groups $P1$ and $P2_1$. This search should also produce crystal structures in $P\bar{1}$ and the experimental space group $P2_1/c$ with 2,2′-bipyridine lying at an inversion center. PLATON[56] was used to detect cases with higher symmetry than that imposed during lattice energy minimization.

Lattice Energy Minimizations. The minimization problem

$$\min_{\mathbf{X},\boldsymbol{\xi}^f} E_{latt} = U_{inter}(\mathbf{X},\boldsymbol{\xi}^f;\boldsymbol{\xi}^r) + \Delta E_{intra}(\boldsymbol{\xi}^f) \qquad (1)$$

is solved at each of the three stages with models that differ in the approximations that they use for the representations of both inter- and intramolecular energy terms and the rigid degrees of freedom $\boldsymbol{\xi}^r$, as summarized in Figure 1. These differences result in a specific formulation for each stage and require separate solution procedures. The methodologies used in the first stage (CrystalPredictor[41,42]) and third stage (DMAflex[29]) have already been presented elsewhere and will therefore be described only briefly here along with the parameters used. The new DMAflex-Quick algorithm used in stage 2 will be discussed in more detail.

Molecular Model and Intramolecular Energy Potential. For both stage 1 and stage 2 minimizations, the values of the rigid degrees of freedom $\boldsymbol{\xi}^r$ are taken from the HF/6−31G(d,p) global conformational minimum and held constant. The intramolecular energy ΔE_{intra} is expressed as a function of the flexible torsions using multidimensional restricted Hermite interpolants[42] over a relatively coarse grid of tabulated quantum mechanical values of

$$\Delta E_{intra}(\boldsymbol{\xi}^f) = \min_{\tilde{\boldsymbol{\xi}},\text{s.t.}\tilde{\boldsymbol{\xi}}^f=\boldsymbol{\xi}^f} E_{intra}(\tilde{\boldsymbol{\xi}}) - E_{intra}^{gas} \qquad (2)$$

and $\partial\Delta E_{intra}/\partial\boldsymbol{\xi}^f$ covering the conformational range shown in Table 1. The constant E_{intra}^{gas} is defined as the global molecular energy minimum $E_{intra}^{gas} \equiv \min_{\boldsymbol{\xi}} E_{intra}(\boldsymbol{\xi})$.

In the third minimization stage, using DMAflex,[29] ΔE_{intra} is expressed as a function of $\boldsymbol{\xi}^f$ using "on-the-fly" quantum mechanical calculations, so that the rigid degrees of freedom, $\boldsymbol{\xi}^r$, are allowed to relax in response to changes in the flexible torsion angles using the same isolated-molecule calculation:

$$\boldsymbol{\xi}^r(\boldsymbol{\xi}^f) = \arg\min_{\tilde{\boldsymbol{\xi}},\text{s.t.}\tilde{\boldsymbol{\xi}}^f=\boldsymbol{\xi}^f} E_{intra}(\tilde{\boldsymbol{\xi}}) \qquad (3)$$

In this work, all quantum mechanical calculations used to optimize molecular structures and compute intramolecular energies were performed at the HF/6−31G(d,p) level of theory using the electronic structure packages GAUSSIAN[57] or GAMESS.[58,59] Hence, the methodology only relies on quantum mechanical estimates of the intramolecular energy $\Delta E_{intra}(\boldsymbol{\xi}^f)$, providing greater accuracy in general than empirical intramolecular force fields.[40,60]

Stage 1: Isotropic Intermolecular Potential (CrystalPredictor). The intermolecular energy is computed with an isotropic atom-atom model. The repulsion−dispersion interactions are modeled with the empirical exp-6 potential developed by Williams.[61] The electrostatic interactions are

	Molecular model	Intramolecular energy model	Intermolecular energy model	
			Electrostatic model	Repulsion-dispersion & other terms
Stage 1 (CrystalPredictor[41,42]) CPU cost: seconds	"Flexible" degrees of freedom (d.o.f.) ξ^f optimized within lattice energy minimization "Rigid" d.o.f. ξ^r frozen to their values in the quantum-mechanically optimized isolated molecule	Approximated with Hermite interpolants on a quantum mechanical (QM) pre-constructed grid	Atomic charges: Computed once for the gas phase global conformational minimum or conformationally dependent using Hermite interpolants on a QM pre-constructed grid	Isotropic, 12-6 or exp-6
Stage 2 (DMAflex-Quick) CPU cost: between 30 min and 3-4 hours			Atomic multipoles up to rank 4: Computed for a reference molecular geometry and rotated analytically with the local environment of the atoms	Isotropic, 12-6 or exp-6 or Any desired improvement that is cost effective in terms of CPU time versus re-ranking, such as use of anisotropic repulsion or C_8 dispersion terms
Stage 3 (DMAflex[29]) CPU cost: between 1 and 3 days	"Flexible" d.o.f. ξ^f optimized within lattice energy minimization "Rigid" d.o.f. ξ^r partially relaxed by isolated-molecule QM optimizations with flexible d.o.f. constrained	Accurate on-the-fly QM calculations	Atomic multipoles up to rank 4: Full conformational dependence with on-the-fly QM calculations of the molecular charge density	CPU intensive improvements such as use of distributed polarizability model

Figure 1. Overview summary of the multistage lattice minimization method. Indicative CPU cost is per crystal structure of the molecules used in this study.

modeled with a set of conformationally invariant atomic charges, fitted to the MP2(fc)/6−31G(d,p) electrostatic potential for the HF/6−31G(d,p) globally optimized molecule in isolation. The electrostatic potential is sampled using GAMESS[58,59] on a dense geodesic mesh of 32 layers, with the first layer at 1.4 times the van der Waals radii and successive layers separated by 0.025 times the van der Waals radii. For consistency with the force field specification,[61] the X−H bond lengths are foreshortened by 0.1 Å for both the fitting of charges and lattice energy minimizations. The repulsion−dispersion interactions are damped to zero at 18 Å using a quintic spline between 15 and 18 Å and summed in direct space. Charge−charge interactions are computed using Ewald summation.[62]

CrystalPredictor minimizes the lattice energy by simultaneously varying both the variables determining the lattice structure and the flexible torsion angles. It uses analytical gradients and a successive quadratic programming algorithm[41,42] that ensures robust and efficient minimization even when the starting point is far away from the minimum structure. Each minimization is very quick, requiring typically no more than a few CPU seconds even for cocrystal structures. In this study, computations were distributed over a 20-node Beowulf cluster. Fifty thousand minimizations were sufficient to provide a reasonable degree of confidence that the search for single-component crystals was complete. However, for the two-component searches, we had to perform significantly more minimizations: 565 000 for the more flexible 4-aminobenzoic acid:4-nitrophenylacetic acid cocrystal and 325 000 for the 4-aminobenzoic acid:2,2′-bipyridine cocrystal in 1:1 stoichiometry (plus 190 000 in 2:1 stoichiometry). Despite the enormous increase in computational effort, we occasionally encountered cocrystal structures within a few kJ mol^{-1} of the global minimum that were located only once, while other minima were much broader[11,40] and were found several tens of times. The stage 1 minima were clustered by comparing intermolecular distances; between 700 and 3000 unique minima were retained for further processing, depending on the system.

Stage 2: Improved Electrostatic Interactions (DMAflex-Quick). In stage 2, a key approximation of the model used in stage 1 minimizations is lifted by replacing the atomic charge model with an anisotropic model for the intermolecular electrostatic interactions comprising distributed atomic multipole moments up to rank 4. In the interest of computational efficiency, the multipole moments can be

computed only for a limited set of reference conformations from the MP2(fc)/6−31G(d,p) isolated-molecule charge density[57] using distributed multipole analysis.[63] Each atom is then assigned its own local axis system using two directly connected atoms (or, in the case of terminal atoms, a directly connected atom and a second neighbor). The reference multipole moments are then converted to their Cartesian form and rotated to the local axis system of each atom. In stage 2, the locally expressed reference multipoles are assumed to be constant despite the changes in flexible torsion angles in the course of lattice energy minimization. Thus, the conformational variability of the electrostatic model is accounted for in a limited manner via the analytical rotation of the local atomic multipoles to the molecular axis system of each newly generated conformation:

$$Q^{i}_{k_1 k_2 \cdots k_n}(\boldsymbol{\xi}^{\mathrm{f}};\boldsymbol{\xi}^{\mathrm{r}}) \approx \sum_{k'_1}\sum_{k'_2}\cdots\sum_{k'_n} R^{i}_{k_1 k'_1}(\boldsymbol{\xi}^{\mathrm{f}};\boldsymbol{\xi}^{\mathrm{r}}) R^{i}_{k_2 k'_2}(\boldsymbol{\xi}^{\mathrm{f}};\boldsymbol{\xi}^{\mathrm{r}})\cdots R^{i}_{k_n k'_n}(\boldsymbol{\xi}^{\mathrm{f}};\boldsymbol{\xi}^{\mathrm{r}}) Q^{i,\mathrm{reference}}_{k'_1 k'_2 \cdots k'_n} \quad (4)$$

where the multipole moment of rank n for atom i is computed by using the rotation matrix $\mathbf{R}(\boldsymbol{\xi}^{\mathrm{f}};\boldsymbol{\xi}^{\mathrm{r}})$ that transforms the local axis system to the molecular one. After each such rotation, the multipoles in the molecular axis system are converted[64] to the spherical-tensor formalism[65] and used to model the intermolecular electrostatic interactions with the program DMACRYS.[66,67]

The choice of reference conformations needs careful consideration depending on the conformational transferability[68] of the multipoles. Calculations on *o*-chlorobenzoic acid[69] show that using a single, gas-phase reference conformation can be a poor approximation, and this may also be the case for other systems that contain a flexible functional group in close proximity to other groups. In such cases, it is necessary to compute the multipole moments for the starting conformation or even to update the local multipole moments after every significant change in molecular geometry.[31] However, all the molecules in this study are para-substituted, so it was assumed sufficient to use the multipole moments of the gas-phase HF/6−31G(d,p) global conformational minimum for all DMAflex-Quick minimizations, knowing that the conformational dependence of the multipoles would be accounted for in stage 3.

The lattice energy minimization is formulated as a two-level optimization problem:

$$\min_{\mathbf{X},\boldsymbol{\xi}^{\mathrm{f}}} E_{\mathrm{latt}} = \min_{\boldsymbol{\xi}^{\mathrm{f}}}\left\{\Delta E_{\mathrm{intra}}(\boldsymbol{\xi}^{\mathrm{f}}) + \min_{\mathbf{X}} U_{\mathrm{inter}}(\mathbf{X};\boldsymbol{\xi}^{\mathrm{f}},\boldsymbol{\xi}^{\mathrm{r}})\right\} \quad (5)$$

with the inner minimization being carried out using the rigid-body lattice energy minimization. The outer minimization problem is solved using a Broyden−Fletcher−Goldfarb−Shanno (BFGS), quasi-Newton minimization algorithm,[70] using numerical gradients of U_{inter} with respect to the flexible degrees of freedom $\boldsymbol{\xi}^{\mathrm{f}}$, that are computed using forward finite differences with a 1.5° perturbation. The gradients $\partial\Delta E_{\mathrm{intra}}/\partial\boldsymbol{\xi}^{\mathrm{f}}$ are computed analytically by differentiating the Hermite interpolants as in the first stage. Charge−charge, charge−dipole, and dipole−dipole interactions are calculated with Ewald summation.[62] Repulsion−dispersion and higher multipole contributions are evaluated in direct space with a 30 Å cutoff to ensure that discontinuities in U_{inter}, due to molecules coming in and out of the cutoff distance do not lead to premature termination of the outer optimization.

At this stage, it would, in principle, be possible to also improve the modeling of the other contributions to the lattice energy, for example by introducing anisotropic atom−atom repulsion[71] or other more theoretically justified model intermolecular potentials.[30,72] However, we did not investigate this option; instead stage 2 minimizations were performed using the empirical, isotropic exp-6 model of Williams[61] but with a modified carboxylic proton−pyridine nitrogen repulsion term as discussed in Supporting Information.

The CPU cost per DMAflex-Quick lattice energy minimization is typically between 30 min and 3−4 h on a modern processor, depending on the number of flexible torsion angles. This timing would increase if conformation-specific initial charge density calculations were necessary. The DMAflex-Quick minima were clustered by comparing the 15-molecule (20-molecule for structures with more than one crystallographically independent molecule) coordination sphere[73] with a threshold of 0.4 Å, and the unique minima were passed to stage 3. To speed up the clustering process, structures that differed by 2 kJ mol^{-1} or more in energy were classed as different without further comparisons.

Stage 3: On-the-Fly Quantum Mechanical Calculations To Derive Conformation-Dependent Multipoles and Intramolecular Energies (DMAflex). The accuracy of lattice energies computed at stage 2 is limited because of three main reasons:

- The conformational transferability of the reference multipole moments is limited.[35,36]
- The rigid degrees of freedom $\boldsymbol{\xi}^{\mathrm{r}}$ are not relaxed in response to changes in the flexible torsions and the intermolecular forces.
- The computation of the intramolecular energy $\Delta E_{\mathrm{intra}}(\boldsymbol{\xi}^{\mathrm{f}})$ and its gradients $\partial\Delta E_{\mathrm{intra}}/\partial\boldsymbol{\xi}^{\mathrm{f}}$ is subject to interpolation errors arising from the use of coarse grids.

Stage 3 minimizations use DMAflex[29] to solve the two-level minimization shown in eq 5. Most sources of stage 2 errors identified above are addressed via the use of "on-the-fly" quantum mechanical calculations to compute the HF/6−31G(d,p) $\Delta E_{\mathrm{intra}}$ and the MP2(fc)/6−31 g(d,p) charge density and distributed multipole moments for the specific conformation in each outer minimization step. It is worth noting, however, that the effect of the intermolecular forces on rigid degrees of freedom $\boldsymbol{\xi}^{\mathrm{r}}$ is still neglected at stage 3, as $\boldsymbol{\xi}^{\mathrm{r}}$ are only relaxed within the isolated-molecule optimization.

We found that most of the large change in lattice energy from stage 2 occurs during the first stage 3 lattice energy evaluation using conformation-specific multipoles. Iterating the stage 3 DMAflex optimization to convergence typically further lowers the lattice energy by no more than 3 kJ mol^{-1} per molecule in the asymmetric unit. Hence, in the interest of computational efficiency, we started stage 3 simply by recalculating the lattice energy of all structures passed from stage 2 using multipole moments and rigid degrees of freedom $\boldsymbol{\xi}^{\mathrm{r}}$ recalculated to be consistent with the values of

Table 2. Lattice Energy Minimization of Experimentally Determined Crystal Structures

	lattice energy partitioning[a] (kJ mol^{-1})			lattice and conformational details							
	E_{latt}	ΔE_{intra}		a (Å)	b (Å)	c (Å)	β (deg)	density (g cm^{-3})	rmsd$_{mol}$[b] (Å)		rmsd$_{cs}$[b](Å)
Single-Component Crystals											
4-Aminobenzoic Acid Polymorph α, $Z' = 2$, $P2_1/n$ (CSD ref AMBNAC01,[43] $R = 7.03\%$, RT)											
experimental		2.57	2.08	18.551	3.860	18.642	93.6	1.367			
stage 1	−92.60	1.25	1.32	18.481	3.760	18.919	93.8	1.388	0.049	0.065	0.179
stage 2	−107.21	0.93	1.28	18.467	3.747	18.788	93.2	1.403	0.058	0.072	0.171
stage 3	−107.15	0.73	0.71	18.399	3.772	18.801	92.7	1.398	0.040	0.051	0.164
4-Aminobenzoic Acid Polymorph β, $Z' = 1$, $P2_1/n$ (CSD ref AMBNAC04,[45] $R = 4.99\%$, RT)											
experimental		2.68		6.278	8.583	12.365	100.1	1.389			
stage 1	−92.84	1.85		6.244	8.582	12.576	100.0	1.372	0.023		0.133
stage 2	−109.78	1.31		6.265	8.551	12.182	99.1	1.413	0.018		0.135
stage 3	−112.74	2.82		6.314	8.441	12.300	99.9	1.411	0.013		0.137
2,2′-Bipyridine, $Z' = 1/2$, $P2_1/n$ (CSD ref BIPYRL04,[98] $R = 2.98\%$, 123 K)[c]											
experimental		0.00		5.486	6.166	11.609	95.3	1.326			
stage 1	−79.61	0.00		5.567	6.349	11.442	96.9	1.291	0.013		0.190
stage 2	−81.41	0.00		5.680	6.180	11.490	95.8	1.293	0.013		0.165
stage 3	−81.41	0.00		5.680	6.180	11.490	95.8	1.293	0.013		0.165
4-Nitrophenylacetic Acid, $Z' = 1$, *Pbca* (CSD ref SEMTAF01, $R = 4.1\%$, 110 K)[99]											
experimental		0.76		14.982	7.010	15.885	−	1.442			
stage 1	−99.41	0.72		15.321	7.014	15.829	−	1.414	0.053		0.388
stage 2	−117.79	1.01		15.175	6.810	16.070	−	1.449	0.047		0.241
stage 3	−117.31	1.10		15.176	6.823	16.023	−	1.450	0.052		0.243
Cocrystals											
4-Aminobenzoic Acid:2,2′-Bipyridine, $Z' = 3/2$, $P2_1/c$ (CSD ref DAQYUQ,[46] $R = 5.73\%$, RT)[c]											
experimental		0.48	0.00	15.820	4.309	16.466	107.2	1.332			
stage 1	−259.94	0.69	0.00	16.443	4.140	16.284	103.1	1.323	0.071	0.077	0.472
stage 2	−310.02	1.33	0.00	16.165	4.264	15.787	106.0	1.367	0.081	0.077	0.295
stage 3	−309.68	1.33	0.00	16.167	4.286	15.721	105.9	1.365	0.079	0.077	0.305
4-Aminobenzoic Acid:4-Nitrophenylacetic Acid, $Z' = 1$, $P2_1/a$ (CSD ref RILJOL,[97] $R = 5.75\%$, RT)											
experimental		3.24	4.94	15.272	5.910	17.194	108.8	1.452			
stage 1	−188.89	0.36	4.14	15.154	5.928	16.933	106.0	1.445	0.024	0.065	0.255
stage 2	−217.48	0.79	5.82	15.276	5.876	16.752	104.4	1.452	0.027	0.042	0.322
stage 3	−217.63	2.62	4.93	15.059	5.922	16.850	105.2	1.458	0.027	0.034	0.251

[a] The HF/6−31G(d,p) intramolecular energy ΔE_{intra} is shown separately for each molecule in the asymmetric unit in the order shown in the system title. The experimental ΔE^{intra} refers to the HF/6−31G(d,p) molecular energy of the gas-phase optimized molecule with the flexible torsion angles identified in Scheme 1 constrained to their experimental values. [b] Hydrogen atoms were omitted for both molecular (rmsd$_{mol}$) and coordination sphere (rmsd$_{cs}$) root-mean-square comparison[73] because of the uncertainties in the X-ray experimental determination of their position. For the calculation of rmsd$_{cs}$ we used 15 and 20 molecules in the coordination sphere for structures with one or more crystallographically independent molecule respectively. The rmsd$_{mol}$ for each molecule in the asymmetric unit is given in the order shown in the system title. [c] Minimizations performed by reducing the symmetry to $P2_1$ to have complete molecules in the asymmetric unit. The minimized crystal retained[56] the higher symmetry of the experimental crystal. E_{latt} for the 4-aminobenzoic acid:2,2′-bipyridine cocrystal refers to 2 mol of 4-aminobenzoic acid and 1 mol of 2,2′-bipyridine.

the flexible degrees of freedom ξ^f determined at stage 2 (a "single-point DMAflex minimization"); we then applied clustering to the resulting structures, and then *fully* minimized the 25 most stable, unique ones among these.

The outer minimization in DMAflex[29] is solved to convergence with a simplex algorithm[70] because values of the $\partial U_{inter}/\partial \xi^f$ gradients are not readily available. An unfortunate consequence of this is that flexible torsion angles do not generally converge to an accuracy greater than 1−2°. The computational cost for a single-point and a full DMAflex minimization varied between 30 min and 2 h, and 1 and 3 days, respectively, depending on the system.

Results

Reproduction of Experimentally Determined Crystals. The overall reproduction quality of the experimental crystal structures with the models used in the three minimization stages is examined in Table 2. All three models reproduce the known crystal structures well, with errors in the coordination sphere[73] (rmsd$_{cs}$) not exceeding 0.5 Å. It is worth noting that the reproduction accuracy achieved with the models used in stages 2 and 3 is generally comparable.

The molecular conformations are also described well, with errors in the non-hydrogen torsion angles not exceeding 10° with the most accurate (DMAflex) model (Table S3 in Supporting Information). The largest discrepancies are observed for the rotation and pyramidalization of the amine group, although these could well be due to experimental uncertainties as well as problems[74] with the *ab initio* calculation of ΔE_{intra}.

Despite the similar overall reproduction quality with the three models, the predicted absolute and relative lattice energies are significantly different. The atomic charge model consistently underestimates the crystal energies compared with the multipole models, but the underestimation varies with the type of molecule, it being larger for hydrogen bonded systems. The stage 3 model predicts that the β polymorph of 4-aminobenzoic acid is 5.6 kJ mol^{-1} more stable than the α one, whereas stage

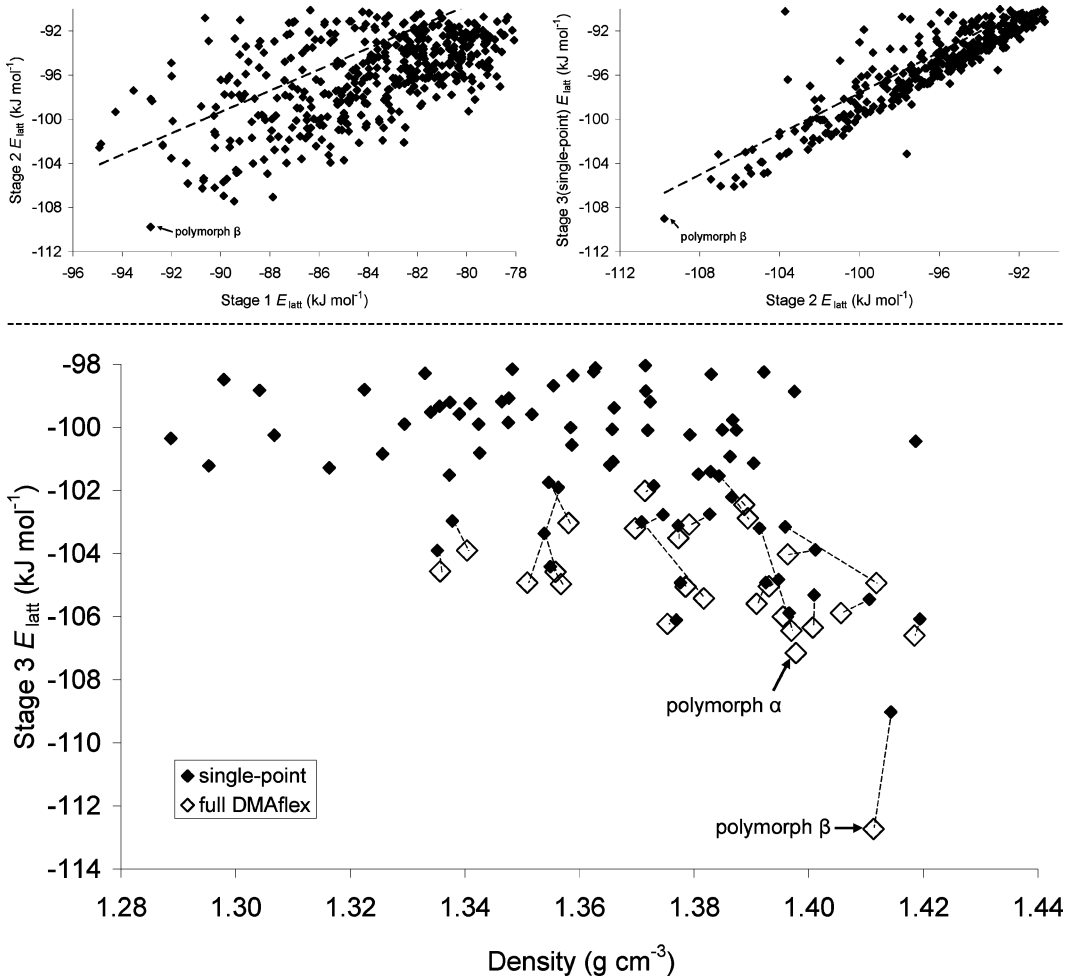


Figure 2. Crystal structure prediction of 4-aminobenzoic acid. The top two figures illustrate the extent of reranking in the three minimization stages; the dashed black lines are linear regressions to guide the eye. The bottom figure shows the lattice energy−density landscape obtained in the third minimization stage with the full and open symbols denoting single-point and full DMAflex minimizations, respectively. The minima obtained when the experimentally determined polymorphs are minimized with the same computational model are also indicated. The experimental $Z' = 2$ polymorph α was fully minimized with DMAflex for completeness, although it was not within the scope of the search. Carrying over the 900 most stable CrystalPredictor structures ($E_{latt,stage1} < -82$ kJ mol^{-1}) to stage 2 was sufficient to locate the majority of structures with single-point DMAflex energies (stage 3 in top right figure) up to 8 kJ mol^{-1} higher than the corresponding global minimum.

1 indicates almost no energy difference between the two polymorphs. The experimental enthalpy difference[75] is 2.5 kJ mol^{-1} at 86 °C and 5 kJ mol^{-1} at 25 °C. Although we cannot confidently extrapolate to 0 K and the modeling of form α only uses a fully ordered structure despite evidence of twinning, disorder, and possibly polytypism,[75] it is certain that the predicted stability difference using multipoles is more realistic compared with the negligible energy difference predicted with atomic charges. We note that the structural reproduction and relative stability of 4-aminobenzoic acid polymorphs using an earlier repulsion−dispersion parametrization[76] in part 1 was significantly worse (see also section S1 in Supporting Information).

With the stage 1 model, both cocrystals are predicted to be marginally (Table 2) less stable than the sum of lattice energies of the most stable polymorphs of the components. With the stage 3 model, the 4-aminobenzoic acid:2,2′-bipyridine cocrystal becomes thermodynamically favorable compared with the most stable single-component polymorphs. However, the 4-aminobenzoic acid:4-nitrophenylacetic acid cocrystal remains unstable, and its stability difference from the single-component crystals increases to 12.4 kJ mol^{-1}.

Crystal Structure Prediction. As shown in Figures 2−6, in all cases the experimentally observed structure was identified as a local lattice energy minimum. This is indicative of the effectiveness of the search algorithm despite the large number of minimization variables.[11] The experimental structures of 4-aminobenzoic acid (Figure 2) and 4-nitrophenylacetic acid (Figure 4) were both found with DMAflex at the corresponding global minima, while the experimental 2,2′-bipyridine crystal structure (Figure 3) was the second most stable structure, only 0.05 kJ mol^{-1} above the global minimum.

Figures 5 and 6 show the crystal structure prediction results for the cocrystals in a form that allows comparison with the pure component lattice energies. The stability of a cocrystal that contains n molecules of A and m molecules of B in the asymmetric unit was expressed as the lattice energy of the

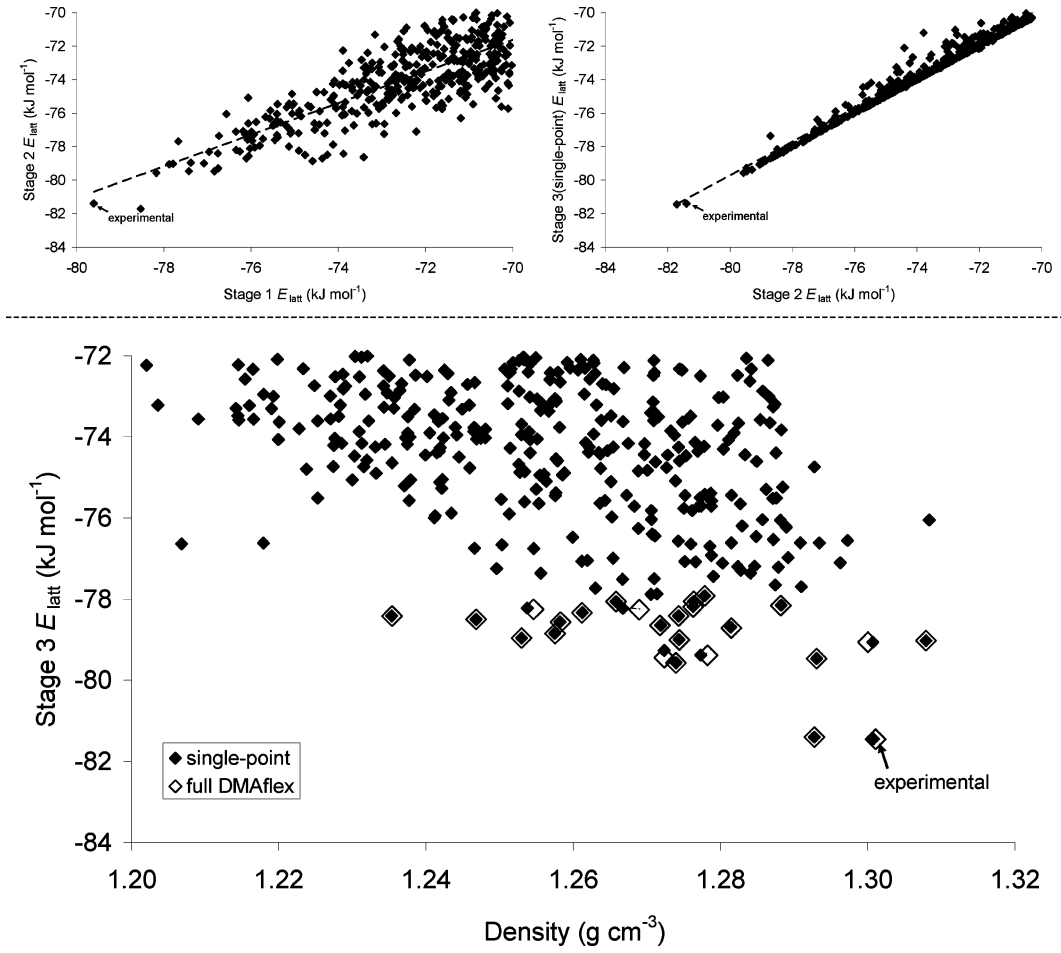


Figure 3. Crystal structure prediction of 2,2′-bipyridine. The top two figures illustrate the extent of reranking in the three minimization stages; the dashed black lines are linear regressions to guide the eye. The bottom figure shows the lattice energy−density landscape obtained in the third minimization stage with the full and open symbols denoting single-point and full DMAflex minimizations, respectively. The minimum obtained when the experimentally determined structure is minimized with the same computational model is also indicated. Carrying over the 700 most stable CrystalPredictor structures ($E_{latt,stage1} < -72$ kJ mol^{-1}) to stage 2 was sufficient to locate the majority of structures with single-point DMAflex energies (stage 3 in top right figure) up to 8 kJ mol^{-1} higher than the corresponding global minimum.

cocrystal (per mol of asymmetric unit) minus n times the predicted global lattice energy minimum of molecule A minus m times the predicted global lattice energy minimum of molecule B. Figure 5 shows that the experimentally determined 4-aminobenzoic acid:2,2′-bipyridine cocrystal was identified as the most stable packing arrangement in the search with 2:1 stoichiometry. However, there is one cocrystal with a 1:1 stoichiometry that is predicted to be more stable than the experimental form.[77] Hence, our predictions do not explain why 4-aminobenzoic acid:2,2′-bipyridine cocrystal adopts a 2:1 stoichiometry, a particularly challenging task given the small energy differences involved.[18] However, as several of the predicted cocrystal structures with both stoichiometries are more stable than the component crystals, the model predicts that cocrystallization is thermodynamically favorable. In contrast, Figure 6 shows that none of the hypothetical 4-aminobenzoic acid:4-nitrophenylacetic acid DMAflex minimized structures are energetically competitive with the DMAflex global lattice energy minima of the components, as all the relative lattice energies are positive. Moreover, the minimized experimental 4-aminobenzoic acid:4-nitrophenylacetic acid cocrystal structure was poorly ranked at 10.4 kJ mol^{-1} above the global minimum, with dozens of structures being more stable, one of which is very similar to the experimental form and 2 kJ mol^{-1} more stable.

The extent of reranking in the three minimization stages is molecule-dependent. The least affected is 2,2′-bipyridine (Figure 3), a molecule which lacks hydrogen bond donors and hence the ability to form strongly directional intermolecular interactions. In this case, the point charge model locates the known crystal structure at the global minimum despite its limitations in describing $\pi-\pi$ interactions,[78] and subsequent minimizations give little change. This contrasts with the energy reranking observed for all other systems, which can involve intermolecular hydrogen bonds. The experimental 4-aminobenzoic acid crystal structure was the fifth most stable in the CrystalPredictor search (Figure 2), approximately 2 kJ mol^{-1} above the global minimum that contains carboxylic dimers. The DMAflex model predicts that the experimental crystal, which contains COOH···NH$_2$ hydrogen bonds, clearly stands out as the unique, most stable packing arrangement separated by over 6 kJ mol^{-1} from the second most stable structure (Figure 2, Table 3). The

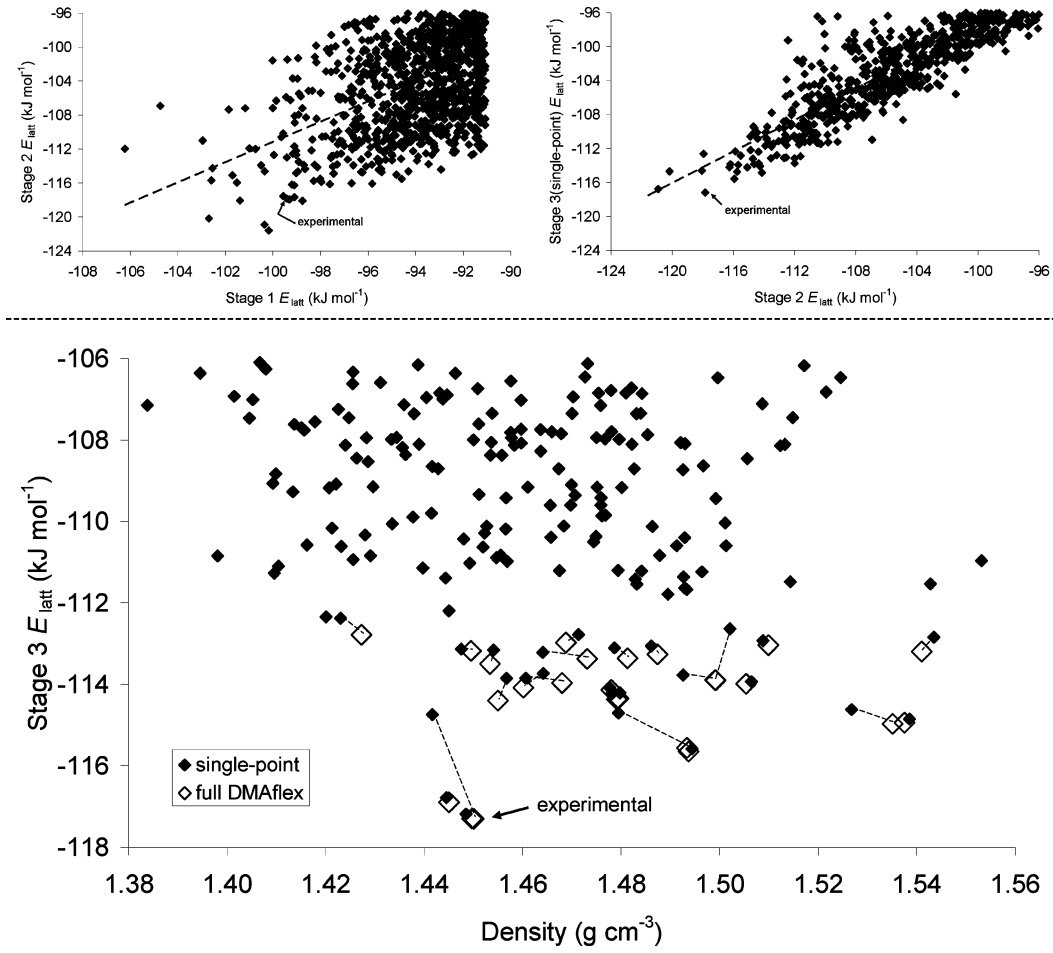


Figure 4. Crystal structure prediction of 4-nitrophenylacetic acid. The top two figures illustrate the extent of reranking in the three minimization stages; the dashed black lines are linear regressions to guide the eye. The bottom figure shows the lattice energy−density landscape obtained in the third minimization stage with the full and open symbols denoting single-point and full DMAflex minimizations, respectively. The structure obtained when the experimentally determined structure is minimized with the same computational model is also indicated. We carried over the 1400 most stable CrystalPredictor structures to stage 2, but the search is unlikely to be complete because of the degree of reranking.

improvements with the computational model are even more evident in the case of 4-nitrophenylacetic acid. Figure 4 shows that the known crystal structure was, at stage 1, poorly ranked and 6 kJ mol^{-1} above the global minimum that contained a catemer hydrogen bond motif. In contrast, the DMAflex model predicts that the experimental structure that contains carboxylic dimers corresponds to the most stable packing arrangement.

Figures 2−6 show that the improvement in the energy ranking of the experimental structures is greater on changing from point charges to distributed multipoles (from stage 1 to stage 2) than in going from analytically rotated to accurate multipoles (from stage 2 to stage 3). Nevertheless, assuming conformational transferability of the intermolecular electrostatic model is also too crude an approximation for the accuracy needed in crystal structure prediction. It is not surprising that the DMAflex-Quick error is greater for 4-nitrophenylacetic acid than 4-aminobenzoic acid, because its structures vary far more in conformation. Several of the hypothetical 2,2′-bipyridine cocrystal structures were also erroneously favored by DMAflex-Quick because they had significantly nonplanar 2,2′-bipyridine conformations, which led to significant errors arising from the analytical rotation of the multipoles of the planar gas-phase geometry. In all cases, the energy lowering caused by completing the DMAflex minimization (indicated by dashed lines in the crystal energy landscapes Figures 2−6) varies in the range of 0−3 kJ mol^{-1} per molecule in the asymmetric unit, which is significantly smaller than the energy changes in the other stages of this multistage improvement in the lattice energy model.

Analysis of Lattice Energy Landscapes. The four most stable predicted crystal structures in each search are given in Table 3 (and the 10 most stable in the Supporting Information), and we have examined the packing arrangement and dominant hydrogen bond motif of the 25 fully DMAflex minimized crystal structures. The majority of 4-aminobenzoic acid hypothetical crystal structures contain carboxylic acid hydrogen bonded dimers as observed in the experimentally determined $Z' = 2$ α form. The chains of OCOH···NH$_2$···OCOH hydrogen bonds found in the β polymorph (at the global minimum) are found only in two other structures. In contrast, all 22 unique, DMAflex-minimized 4-nitrophenylacetic acid crystal structures contain carboxylic acid dimers. In practically all the low energy 2,2′-

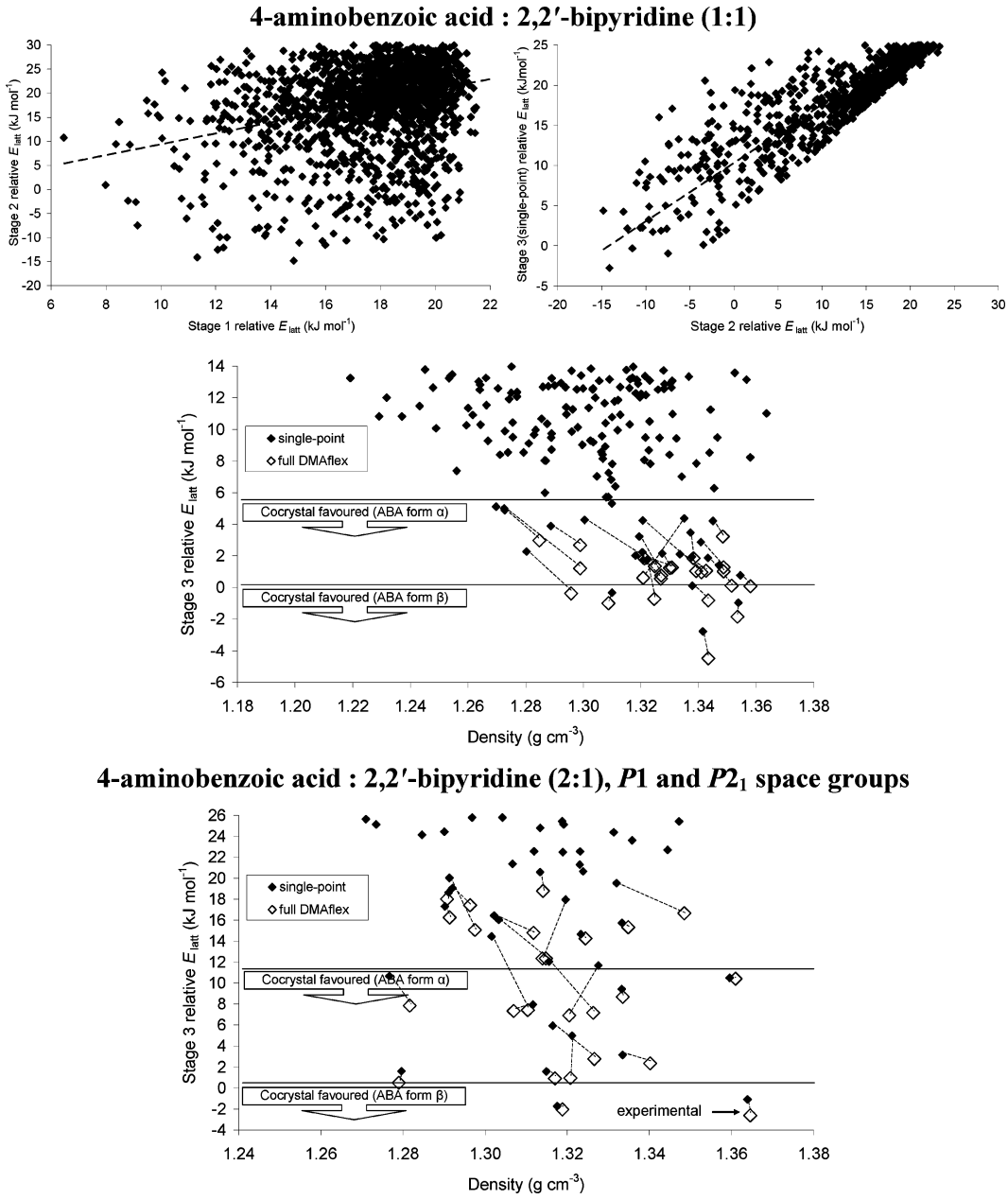


Figure 5. Crystal structure prediction of 4-aminobenzoic acid (ABA):2,2′-bipyridine cocrystal. All lattice energies are reported relative to the sum of the global lattice energy minima of the components with the same computational model. The top two figures illustrate the extent of reranking in the three minimization stages for 1:1 stoichiometry; the dashed black lines are linear regressions to guide the eye. The middle (1:1 stoichiometry) and bottom (2:1 stoichiometry) figures show the lattice energy−density landscape obtained in the third minimization stage with the full and open symbols denoting single-point and full DMAflex minimizations, respectively. The structure obtained when the experimentally determined structure is minimized with the same computational model is also indicated. We carried over the 2000 (1:1) and 750 (2:1) most stable CrystalPredictor structures to stage 2, but the search is unlikely to be complete because of the degree of reranking.

bipyridine structures, the pyridine rings are coplanar, showing that deforming the molecule from this gas-phase, planar conformation does not improve the packing. The hypothetical structure at the global minimum exhibits a tilted edge-to-face (T-shaped) arrangement of the molecular planes and better overlap in the $\pi-\pi$ stacking than the slightly less stable experimental structure.

In the majority of the low energy 4-aminobenzoic acid: 2,2′-bipyridine cocrystal structures, all carboxylic acid protons are hydrogen bonded to pyridine nitrogen atoms; in the case of 1:1 stoichiometry, the second pyridine nitrogen competes with the carboxylic carbonyl for the amine hydrogen atoms. Hence, the computational model correctly predicts that the presence of the pyridine groups disrupts the carboxylic dimers and $COOH\cdots NH_2$ hydrogen bonds of 4-aminobenzoic acid that is consistent with the strong tendency of pyridyl nitrogen atoms to hydrogen bond to carboxyl acid donors.[79−81] In the majority of the predicted stable 1:1 cocrystals, 2,2′-bipyridine adopts a nonplanar conformation, whereas it is planar and on an inversion center in several of the 2:1 cocrystals including the experimental form. The rest of the 2:1 cocrystal structures also have both

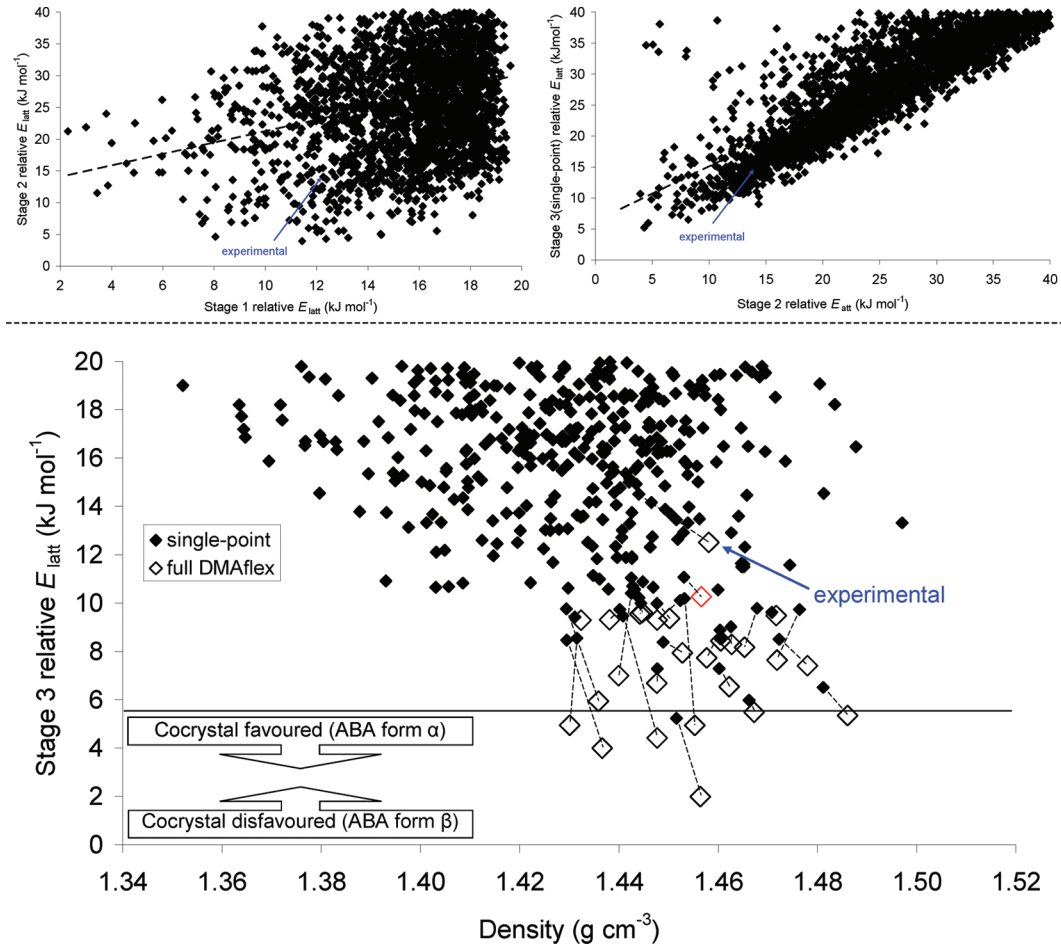


Figure 6. Crystal structure prediction of 4-aminobenzoic acid (ABA):4-nitrophenylacetic acid (1:1) cocrystals. All crystal energies are reported relative to the sum of the global lattice energy minima of the components with the same computational model. The top two figures illustrate the extent of reranking in the three minimization stages; the dashed black lines are linear regressions to guide the eye. The bottom figure shows the lattice energy−density landscape obtained in the third minimization stage with the full and open symbols denoting single-point and full DMAflex minimizations respectively. The structure obtained when the experimentally determined structure is minimized with the same computational model is also indicated; the red diamond shows a hypothetical structure which is similar to the experimental ($rmsd_{cs}$ = 0.34 Å). We carried over the 3000 most stable CrystalPredictor structures to stage 2, but the search is unlikely to be complete because of the degree of reranking.

pyridine nitrogen atoms hydrogen bonded to two carboxylic protons and hence may transform to a slightly more stable symmetric packing if thermal motion were included.

The 4-aminobenzoic acid:4-nitrophenylacetic acid lattice energy landscape exhibits a plurality of hydrogen bonding patterns. A large proportion of the most stable structures contain 4-aminobenzoic acid homodimers. In these structures, the carboxylic acid hydrogen atom of 4-nitrophenylacetic acid is hydrogen bonded to the amine nitrogen; hydrogen bonds are also formed between the amine hydrogen atoms and the carbonyl and nitro groups of 4-nitrophenylacetic acid. The next most dominant hydrogen bond motif is a carboxylic acid heterodimer, with the most stable such structure being found only 3 kJ mol^{-1} above the global minimum. This motif appears in the experimental structure and in 33% of all crystal structures containing two distinct carboxylic acids.[82] There are a limited number of low energy structures in which all hydrogen bond acceptors and donors are associated in chains forming intricate 3-D patterns. There is also a noticeable absence of stable structures containing 4-nitrophenylacetic acid homodimers. The CSD contains only five cocrystals of 4-aminobenzoic acid with carboxylic acids, three of which (including the cocrystal with 4-nitrophenylacetic acid) contain carboxylic acid heterodimers. In the cocrystal with nicotinic acid and the ternary cocrystal with 2,4,6-trinitrobenzoic acid and 1,3,5-trinitrobenzene, 4-aminobenzoic acid does form the homodimers seen in the low energy predicted structures (Table 3). The stronger tendency of 4-aminobenzoic acid to associate in $R_2^2(8)$ carboxylic acid dimers compared with 4-nitrophenylacetic acid is consistent with the differences in electrostatic potential and overall spatial arrangement of the carboxylic groups (see Figure S4 in Supporting Information).

Discussion

Based on our results, the structure generation in stage 1 appears to be sufficiently extensive to locate the experimental crystal structures, including the challenging 4-aminobenzoic acid:2,2′-bipyridine cocrystal that required a search with three crystallographically independent molecules. However, the energy rank of the experimental structures at stage 1 is

Table 3. Most Stable DMAflex Fully-Minimized Crystal Structures[a]

rank, space group[b]	lattice energy partitioning[c] (kJ mol^{-1}) E_{latt}	ΔE_{intra}	conventional lattice cell[d] a (Å)	b (Å)	c (Å)	α, β, γ (deg)	density (g cm^{-3})	packing/dominant hydrogen bond motif
			Single-Component Crystals					
			4-Aminobenzoic Acid					
1, $P2_1/c$	**−112.74**	**2.82**	**6.336**	**8.445**	**12.246**	**100.0**	**1.411**	**OCOH···NH2···OCOH four member rings**
2, $P2_1/c$	−106.60	0.39	3.838	15.271	10.966	92.3	1.419	carboxylic dimers
3, $P2_1$	−106.44	6.12	4.009	14.489	5.661	97.5	1.397	OCOH···NH2···OCOH chains
4, $P2_1/c$	−106.35	0.50	3.707	11.100	15.979	98.5	1.401	carboxylic dimers
			2,2′-Bipyridine					
1, $P2_1/c$	−81.46	0.05	6.017	11.249	12.951	114.6	1.301	T-arrangement, π-π stacking
2, $P2_1/c$, $Z' = 1/2$	**−81.41**	**0.00**	**5.680**	**6.180**	**12.289**	**111.5**	**1.293**	**limited π-π stacking**
3, $Pna2_1$	−79.57	0.00	13.701	5.061	11.744	–	1.274	limited π-π stacking
4, $P2_1/c$, $Z' = 1/2$	−79.47	0.00	3.787	9.735	10.902	93.4	1.293	sheets, limited π-π stacking
			4-Nitrophenylacetic Acid					
1, *Pbca*	−117.30	1.02	15.178	6.822	16.027	–	1.450	carboxylic dimers
2, $P2_1/c$	−116.90	0.82	8.021	15.254	6.805	90.4	1.445	carboxylic dimers
3, $P2_1/c$	−115.65	1.22	6.511	4.793	26.093	98.4	1.494	carboxylic dimers
4, $P2_1/c$	−114.97	2.30	4.755	15.258	10.923	98.5	1.535	carboxylic dimers
			Cocrystals					
			4-Aminobenzoic Acid:2,2′-Bipyridine (1:1 stoichiometry)					
1, $P2_1/c$	−198.68	2.55, 1.34	3.830	13.066	29.100	95.2	1.343	OCOH···N (pyr), NH2···OCOH
2, $P2_1$	−196.04	2.00, 1.35	3.955	12.623	14.512	96.5	1.354	OCOH···N (pyr), NH2···OCOH
3, $P2_1/c$	−195.20	6.66, 1.31	7.041	30.514	7.341	109.6	1.311	OCOH···N (pyr), NH2···OCOH
4, $P2_1$	−195.01	0.21, 1.34	3.856	13.283	14.217	95.2	1.343	OCOH···N (pyr), NH2···OCOH
			4-Aminobenzoic Acid:2,2′-Bipyridine (2:1 stoichiometry)					
1, $P2_1/c$, $Z' = 3/2$	**−309.55**	**1.32, 0.00**	**16.173**	**4.282**	**15.732**	**106.0**	**1.365**	**OCOH···N (pyr), NH2···OCOH·NH2··NH2**
2, $P2_1/c$, $Z' = 3/2$	−308.98	2.12, 0.00	3.904	13.082	21.311	95.2	1.319	OCOH···N (pyr) NH2···OCOH
3, $P2_1/c$, $Z' = 3/2$	−306.44	4.17, 0.00	8.430	4.689	28.583	98.4	1.279	OCOH···N (pyr), NH2···OCOH·NH2··NH2
4, $P2_1/c$, $Z' = 3/2$	−306.02	1.25, 0.00	3.778	21.677	13.509	101.1	1.317	OCOH···N (pyr), NH2···OCOH
			4-Aminobenzoic Acid:4-Nitrophenylacetic Acid (1:1)[e]					
1, $P2_1/c$	−228.05	1.98, 2.72	7.318	15.287	13.954	111.6	1.456	4-aminobenzoic carboxylic homodimers, OCOH(4-nit)···NH2, NH2···O2N
2, $P\bar{1}$	−226.04	3.99, 1.43	5.102	12.082	13.119	105.8, 101.2 101.4	1.437	4-aminobenzoic carboxylic homodimers, OCOH(4-nit)···NH2, NH2···O2N
3, $P2_1/c$	−225.62	4.42, 2.50	13.166	7.471	17.660	122.8	1.448	4-aminobenzoic carboxylic homodimers, OCOH(4-nit)···NH2, HOCO(4-amn,4-nit)···H2N
4, $P2_1/c$	−225.10	4.93, 1.89	4.496	25.887	12.482	90.26	1.455	4-aminobenzoic carboxylic homodimers, OCOH(4-nit)···NH2, HOCO(4-nit)···H2N, NO2···H2N

[a] Rows in Bold Correspond to the Experimental Structures. [b] Z' is given when 2,2′-bipyridine lies at a special position. [c] The HF/6−31G(d,p) intramolecular energy ΔE_{intra} is shown separately for each molecule in the asymmetric unit in the order shown in the system title. [d] Only cell angles not constrained by symmetry given. [e] The DMAflex-minimized experimental crystal lies 10.4 kJ mol^{-1} above the global minimum.

generally poor. The only exception is 2,2′-bipyridine, the only molecule in this study that lacks hydrogen bond donors. The distributed multipole model drastically changes the energy ordering of the hypothetical structures, resulting in a better ranking of the experimental structures, often without any significant structural changes (cf. Table 2). This emphasizes that the reproduction of known crystal structures by lattice energy minimization is a necessary but by no means sufficient condition for the accuracy of the computational model for the lattice energy.

Figures 2−6 show that the improvement seen by replacing point charges with analytically rotated multipoles is larger than the improvement gained by the use of conformationally dependent multipoles. Hence, the use of DMAflex-Quick followed by a single-point DMAflex minimization is a worthwhile procedure to efficiently locate promising hypothetical crystal structures to perform computationally demanding calculations. In principle, stage 1 structures should be passed in batches of increasing energy to later stages until we generate no additional, distinct minima within a predefined range from the global minimum in the final stage. In practice, if the target was to identify the majority of local minima within 8 and 12 kJ mol^{-1} from the global minimum in the final stage for single- and multicomponent crystals respectively, the objective was achieved only for 4-aminobenzoic acid and 2,2′-bipyridine (Figure 2−3), where we were able to minimize sufficiently large batches of stage 1 structures. The overall energy reranking between stage 1 and the fully DMAflex minimized structures is so severe that all other searches were terminated because of limited computer resources (Figure 4−6). Hence, carrying over several thousand crystal structures to subsequent stages does not guar-

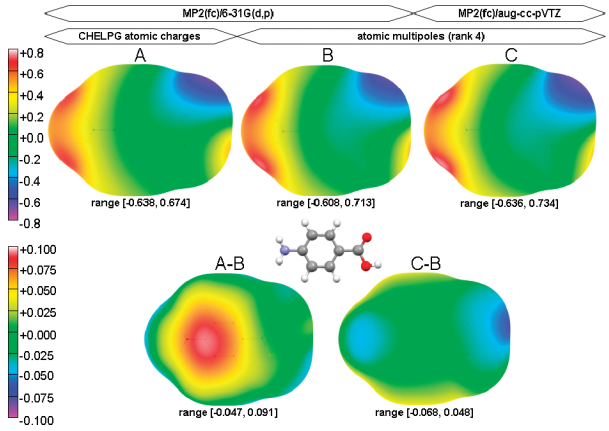


Figure 7. Electrostatic potential (in Volts) (top row) of the HF/6−31G(d,p) globally optimized molecular conformation of 4-aminobenzoic acid on twice the van der Waals radii surface (van der Waals radius for polar hydrogen atoms set to 1 au; radii for other atoms taken from Bondi[94]) computed with the MP2(fc)/6−31G(d,p) atomic charges fitted to the electrostatic potential[95] (A); the DMA[63] atomic multipole moments up to rank 4 derived from the same MP2(fc)/6−31G(d,p) charge density (B) and derived from the MP2(fc)/aug-cc-pVTZ isolated-molecule charge density (C). The bottom row shows electrostatic potential differences A-B and C−B. The maps are displayed using ORIENT[96] for the molecular orientation shown in ball-and-stick.

antee that all low-energy minima are found, even if we assume that the stage 1 search is complete despite its stochastic character.

The reranking from stage 1 to stage 2 is consistent with the large difference in the electrostatic potential around 4-aminobenzoic acid (Figure 7), calculated from an atomic charge and atomic multipole representation of the MP2(fc)/6−31G(d,p) charge density. However, Figure 7 (and also Figure S1 and Table S1 in Supporting Information) shows that increasing the basis set can also have a significant effect, which decreases the predicted stability of 4-aminobenzoic acid structures based on COOH···NH_2 hydrogen bonds relative to those that contain carboxylic dimers. The charge density quality has a smaller impact on the relative stability of 4-nitrophenylacetic acid structures, presumably because all of these contain carboxylic dimers. Thus, the monomer charge-density quality should be carefully considered, particularly when the lower energy structures differ in the hydrogen bonding motif.

This study has relied on empirical, transferable models for the repulsion−dispersion interactions. Part I of this paper showed that the FIT[76] potential severely overestimates the stability difference of 4-aminobenzoic acid polymorphs. This is improved (Table 2) by using the more recent Williams[61] parametrization in this paper, which contains more atom types. However, the Williams potential in conjunction[33,61] with multipole moments underestimates hydrogen bond lengths. This underestimation is particularly severe when the crystal contains COOH···N(pyr) close contacts and hence the modeling of 4-aminobenzoic acid:2,2′-bipyridine cocrystals required a modification of the Williams carboxylic proton - pyridine nitrogen repulsion potential as discussed in Supporting Information. Hence, there is also a noticeable dependence of the computed lattice energy landscape on the repulsion−dispersion parametrization, even when the electrostatic contributions due to strong hydrogen bonding dominate the lattice energy.

Empirically fitted potentials partially absorb the errors in modeling the electrostatic forces and intermolecular polarization effects. Nevertheless, induction is anisotropic and depends strongly on the molecular association in the lattice, which is consistent with the significant difference that the explicit modeling of induction[32] can make in the relative stability of polymorphs with different hydrogen bonding motifs.[53] The hydrogen bonding motif in cocrystals will generally be different from the motif in the component crystal structures; hence, neglecting explicit induction will not cancel out in predicting the thermodynamic grounds for cocrystal formation. The development of molecule-specific models for the repulsion−dispersion[30,72,83] and polarization energies[32] are worthwhile improvements that will complement the accurate representation of the isolated-molecule charge density using distributed multipoles.[33] An advantage of the new search methodology is that it can use such theoretically justified models[30] for calculating the lattice energy at the final stage. An alternative or subsequent improvement in the estimated relative thermodynamic stability would be to investigate the implicit assumption in this study that the differences in vibrational contributions to the free energy are small compared with the lattice energy differences. This assumption is questionable given the density variation of up to 10% in the low energy structures (Figures 2, 4−6).

Despite these limitations, four out of five searches resulted in the experimental crystal structures being very close to the global minimum. The exception is 4-aminobenzoic acid:4-nitrophenylacetic acid cocrystal where the search produced several structures based on 4-aminobenzoic acid homodimers that are more stable than the experimental form. Although this may reflect errors in the relative thermodynamic stability model, it could be a kinetic factor, as self-association of 4-aminobenzoic acid in homodimers has been observed in cocrystals with carboxylic acids.[84] Thus, we cannot exclude the possibility that a homodimer-based polymorph could be found in the future, despite most sets of polymorphic cocrystals[3] in the CSD only differing in the spatial arrangement of identical hydrogen bonded motifs,[85] although exceptions to this generalization do exist.[86] This ability of a computational search to generate alternative hydrogen-bonding motifs makes it a complement to the synthon approach to the design of cocrystals,[87] which relies on the persistence of common hydrogen bonding motifs. On the other hand, cocrystallization of 2,2′-bipyridine with 4-aminobenzoic acid is predicted to almost certainly lead to the disruption of the carboxylic dimers and COOH···NH_2 hydrogen bonds and the formation of carboxylic···N_{arom} hydrogen bonds, in accord with the synthon approach and experimental findings.[80,81]

The proposed methodology allows the efficient identification of stable minima on complex lattice energy surfaces due to the presence of multiple, flexible molecules (or ions) in the asymmetric unit. This knowledge of the lattice energy landscape may prove useful in solving the crystal structures

from powder X-ray data,[88,89] which is often necessary as many cocrystals are synthesized by solid-state grinding. Moreover, the lattice energy landscape can enhance the understanding of the range of possible crystal structures and help to anticipate possible crystallization problems such as disorder and polymorphism.[90] However, the crystallization of cocrystals is also subject to kinetic effects,[91] and hence it is not clear whether all molecular associations generated in the search are really in competition in the nucleation stage of practical cocrystallization processes.

Conclusions

This study presents an extensive, yet efficient, crystal structure prediction methodology to locate thermodynamically stable crystal structures on complex lattice energy surfaces that are typical of multicomponent, flexible systems using conformationally dependent multipole moments and quantum mechanically computed conformational energies. All searches for single- and two-component crystals successfully located the experimentally determined crystal structure and in four out of five cases this lay at, or very close to, the global lattice energy minimum. The ability to predict the structure and thermodynamic stability of cocrystals allows the comparison of the stability of possible cocrystal packing arrangements relative to the single-component crystals. Such applications can greatly enhance our understanding of cocrystal formation, which, at present, relies on a qualitative hierarchy of a small set of chemically intuitive synthons while ignoring the much larger set of secondary, but collectively significant, intermolecular interactions.[92] However, as in part 1 of this paper, it is clear that the energetic advantage of cocrystal formation is small[18] compared with the errors in the model to compute the lattice energies. Nevertheless, this multistage search methodology can be extended to use even more accurate and computationally intensive lattice energy minimization methods[27,93] on a small set of crystal structures that have been found to be energetically competitive.

Acknowledgment. Funding to the Molecular Systems Engineering group from the Engineering and Physical Sciences Research Council (EPSRC) of the UK [EP/E016340] is gratefully acknowledged. Part of this research was carried out within the auspices of the CPOSS project (http://www.cposs.org.uk), funded by the Basic Technology Programme of the Research Councils UK. G.S.K. is acknowledged for funding NI. The assistance of Dr. L. S. Price in preparing this manuscript is gratefully acknowledged.

Supporting Information Available: Sensitivity of the predictions to the model for the intermolecular forces, modification of the Williams repulsion−dispersion parametrization for the 4-aminobenzoic acid:2,2′-bipyridine cocrystal structure prediction, reproduction of flexible torsion angles and hydrogen bond geometries during lattice energy minimization of the experimentally determined crystal structures and tables with the 10 most stable predicted crystal structures for each search. All stage three crystal structures are stored on CCLRC e-Science Centre data portal and are available from the authors on request. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Almarsson, O.; Zaworotko, M. J. *Chem. Commun.* **2004**, *17*, 1889–1896.

(2) Aakeroy, C. B.; Salmon, D. J. *CrystEngComm* **2005**, *7*, 439–448.

(3) Vishweshwar, P.; McMahon, J. A.; Bis, J. A.; Zaworotko, M. J. *J. Pharm. Sci.* **2006**, *95*, 499–516.

(4) Etter, M. C. *Acc. Chem. Res.* **1990**, *23*, 120–126.

(5) Issa, N.; Karamertzanis, P. G.; Welch, G. W. A.; Price, S. L. *Cryst. Growth Des.* **2009**, *9*, 442–453.

(6) Day, G. M.; Cooper, T. G.; Cruz Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A.; Schweizer, B Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem. 2009, in press.

(7) Leusen, F. J. J. *Cryst. Growth Des.* **2003**, *3*, 189–192.

(8) Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. B* **2005**, *109*, 17134–17150.

(9) Cabeza, A. J. C.; Day, G. M.; Motherwell, W. D. S.; Jones, W. *J. Am. Chem. Soc.* **2006**, *128*, 14466–14467.

(10) Hulme, A. T.; Price, S. L. *J. Chem. Theory Comput.* **2007**, *3*, 1597–1608.

(11) van Eijck, B. P.; Kroon, J. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2000**, *56*, 535–542.

(12) Collins, A.; Cooper, R. I.; Watkin, D. J. *J. Appl. Crystallogr.* **2006**, *39*, 842–849.

(13) Collins, A. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2006**, *62*, 897–911.

(14) Desiraju, G. R. *CrystEngComm* **2007**, *9*, 91–92.

(15) Gavezzotti, A. *CrystEngComm* **2008**, *10*, 389–398.

(16) Mohamed, S.; Barnett, S. A.; Tocher, D. A.; Shankland, K.; Leech, C. K.; Price, S. L. *CrystEngComm* **2008**, *10*, 399–404.

(17) Steed, J. W. *CrystEngComm* **2003**, *5*, 169–179.

(18) Cruz-Cabeza, A. J.; Day, G. M.; Jones, W. *Chem. Eur. J.* **2008**, *14*, 8830–8836.

(19) Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2000**, *56*, 697–714.

(20) Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2002**, *58*, 647–661.

(21) Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.;

Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2005**, *61*, 511–527.

(22) Rodriguez-Spong, B.; Price, C. P.; Jayasankar, A.; Matzger, A. J.; Rodriguez-Hornedo, N. *Adv. Drug Delivery Rev.* **2004**, *56*, 241–274.

(23) Day, G. M.; Price, S. L.; Leslie, M. *J. Phys. Chem. B* **2003**, *107*, 10919–10933.

(24) van Eijck, B. P. *J. Comput. Chem.* **2001**, *22*, 816–826.

(25) van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. *J. Phys. Chem. B* **2001**, *105*, 10573–10578.

(26) Aubrey-Medendorp, C.; Swadley, M. J.; Li, T. L. *Pharm. Res.* **2008**, *25*, 953–959.

(27) Neumann, M. A.; Perrin, M. A. *J. Phys. Chem. B* **2005**, *109*, 15531–15541.

(28) Neumann, M. A.; Leusen, F. J. J.; Kendrick, J. *Angew. Chem., Int. Ed.* **2008**, *47*, 2427–2430.

(29) Karamertzanis, P. G.; Price, S. L. *J. Chem. Theory Comput.* **2006**, *2*, 1184–1199.

(30) Misquitta, A. J.; Welch, G. W. A.; Stone, A. J.; Price, S. L. *Chem. Phys. Lett.* **2008**, *456*, 105–109.

(31) van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. *J. Comput. Chem.* **2001**, *22*, 805–815.

(32) Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 522–532.

(33) Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2005**, *5*, 1023–1033.

(34) Cabeza, A. J. C.; Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2006**, *6*, 1858–1866.

(35) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem. Phys. Lett.* **1995**, *238*, 253–260.

(36) Koch, U.; Stone, A. J *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 1701–1708.

(37) Mannfors, B. E.; Mirkin, N. G.; Palmo, K.; Krimm, S. *J. Phys. Chem. A* **2003**, *107*, 1825–1832.

(38) van Eijck, B. P. *J. Comput. Chem.* **2002**, *23*, 456–462.

(39) Nowell, H.; Price, S. L. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2005**, *61*, 558–568.

(40) Day, G. M.; Motherwell, W. D. S.; Jones, W. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1693–1704.

(41) Karamertzanis, P. G.; Pantelides, C. C. *J. Comput. Chem.* **2005**, *26*, 304–324.

(42) Karamertzanis, P. G.; Pantelides, C. C. *Mol. Phys.* **2007**, *105*, 273–291.

(43) Lai, T. F.; Marsh, R. E. *Acta Crystallogr.* **1967**, *22*, 885–893.

(44) Alleaume, M.; Salascim, G.; Decap, J. *C. R. Hebd. Seances Acad. Sci. Ser. C* **1966**, *262*, 416–417.

(45) Gracin, S.; Fischer, A. *Acta Crystallogr. Sect. E: Struct. Rep. Online* **2005**, *61*, O1242−O1244.

(46) Bowers, J. R.; Hopkins, G. W.; Yap, G. P. A.; Wheeler, K. A. *Cryst. Growth Des.* **2005**, *5*, 727–736.

(47) Price, S. L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1996–2009.

(48) Aakeroy, C. B.; Beatty, A. M.; Helfrich, B. A.; Nieuwenhuyzen, M. *Cryst. Growth Des.* **2003**, *3*, 159–165.

(49) Allen, F. H.; Harris, S. E.; Taylor, R. *J. Comput. -Aided Mol. Des.* **1996**, *10*, 247–254.

(50) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. *J. Chem. Inf. Model.* **2008**, *48*, 1–24.

(51) Weng, Z. F.; Motherwell, W. D. S.; Allen, F. H.; Cole, J. M. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2008**, *64*, 348–362.

(52) Cooper, T. G.; Jones, W.; Motherwell, W. D. S.; Day, G. M. *CrystEngComm* **2007**, *9*, 595–602.

(53) Karamertzanis, P. G.; Day, G. M.; Welch, G. W. A.; Kendrick, J.; Leusen, F. J. J.; Neumann, M. A.; Price, S. L. *J. Chem. Phys.* **2008**, *128*, art-244708.

(54) Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2002**, *58*, 380–388.

(55) Sobol', I. M. *USSR Comp. Math. Math+* **1967**, *7*, 86–112.

(56) PLATON A Multipurpose Crystallographic Tool, Spek, A. L. Utrecht University: Utrecht, The Netherlands, 2003.

(57) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J.; 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.; 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.

(58) Schmidt, M. W.; Baldridge, 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.

(59) Gordon, M. S.; Schmidt, M. W. Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry: the first forty years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 1167−1189.

(60) Brodersen, S.; Wilke, S.; Leusen, F. J. J.; Engel, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4923–4931.

(61) Williams, D. E. *J. Comput. Chem.* **2001**, *22*, 1154–1166.

(62) Ewald, P. *Ann. Phys.* **1921**, *64*, 253–287.

(63) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

(64) Stone, A. J. *The Theory of Intermolecular Forces*, 1st ed.; Clarendon Press: Oxford, 1996.

(65) This multipole orientation method was checked against the spherical tensor based manipulations in ORIENT.

(66) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. *J. Comput. Chem.* **1995**, *16*, 628–647.

(67) Welch, G. W. A.; Karamertzanis, P. G.; Price, S. L; Leslie, M. DMACRYS, version 1.05, is a substantial revision of DMAREL.http://www.chem.ucl.ac.uk/cposs/dmacrys/2008.

(68) Price, S. L.; Stone, A. J *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 1755–1763.

(69) Polito, M.; D'Oria, E.; Maini, L.; Karamertzanis, P. G.; Grepioni, F.; Braga, D.; Price, S. L. *CrystEngComm* **2008**, *10*, 1848–1854.

(70) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN 90*; Cambridge University Press: Cambridge, UK, 1996; Vol. 2.

(71) Day, G. M.; Price, S. L. *J. Am. Chem. Soc.* **2003**, *125*, 16434–16443.

(72) Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. *J. Phys. Chem. A* **1999**, *103*, 9883–9890.

(73) Chisholm, J. A.; Motherwell, S. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.

(74) Wang, S. Y.; Schaefer, H. F. *J. Chem. Phys.* **2006**, *124*, art-044303.

(75) Gracin, S.; Rasmuson, A. C. *Cryst. Growth Des.* **2004**, *4*, 1013–1023.

(76) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

(77) The lowest energy packing configuration that consists of two molecules of 4-aminobenzoic acid per one molecule of 2,2′-bipyridine corresponds to the 1:1 cocrystal global minimum crystallizing concomitantly with the 4-aminobenzoic acid global minimum. This packing configuration is 4.5 and 1.9 kJ mol^{-1} more stable compared with the three molecules crystallizing on their own and the global 2:1 cocrystal global minimum, respectively.

(78) Price, S. L *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 2997–3008.

(79) Steiner, T. *Acta Crystallogr., Sect. B* **2001**, *57*, 103–106.

(80) Aakeroy, C. B.; Hussain, I.; Forbes, S.; Desper, J. *CrystEngComm* **2007**, *9*, 46–54.

(81) Shattock, T. R.; Arora, K. K.; Vishweshwar, P.; Zaworotko, M. J. *Cryst. Growth Des.* **2008**, *8*, 4533–4545.

(82) Allen, F. H.; Motherwell, W. D. S.; Raithby, P. R.; Shields, G. P.; Taylor, R *New J. Chem.* **1999**, *23*, 25–34.

(83) 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.

(84) Jebas, S. R.; Balasubramanian, T. *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2006**, *62*, o5621−o5622.

(85) Vishweshwar, P.; McMahon, J. A.; Zaworotko, M. J. Crystal Engineering of Pharmaceutical Co-Crystals. In *Frontiers in Crystal Engineering*, Tiekink, E. R. T., Vittal, J. J., Eds.; John Wiley & Sons Ltd.: Chichester, UK, 2006; pp 25−49.

(86) Sreekanth, B. R.; Vishweshwar, P.; Vyas, K. *Chem. Commun.* **2007**, 2375–2377.

(87) Friscic, T.; Jones, W. *Faraday Discuss.* **2007**, *136*, 167–178.

(88) Tremayne, M.; Grice, L.; Pyatt, J. C.; Seaton, C. C.; Kariuki, B. M.; Tsui, H. H. Y.; Price, S. L.; Cherryman, J. C. *J. Am. Chem. Soc.* **2004**, *126*, 7071–7081.

(89) Schmidt, M. U.; Buchsbaum, C. *Z. Kristallogr.* **2008**, *223*, 418–423.

(90) Copley, R. C. B.; Barnett, S. A.; Karamertzanis, P. G.; Harris, K. D. M.; Kariuki, B. M.; Xu, M. C.; Nickels, E. A.; Lancaster, R. W.; Price, S. L. *Cryst. Growth Des.* **2008**, *8*, 3474–3481.

(91) Desiraju, G. R. *Nat. Mater.* **2002**, *1*, 77–79.

(92) Aakeroy, C. B.; Desper, J.; Elisabeth, E.; Helfrich, B. A.; Levin, B.; Urbina, J. F. *Z. Kristallogr.* **2005**, *220*, 325–332.

(93) Civalleri, B.; Zicovich-Wilson, C. M.; Valenzano, L.; Ugliengo, P. *CrystEngComm* **2008**, *10*, 405–410.

(94) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

(95) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.

(96) Stone, A. J.; Dullweber, A.; Engkvist, O.; Fraschini, E. Hodges, M. P.; Meredith, A. W.; Nutt, D. R.; Popelier, P. L. A.; Wales, D. J. ORIENT: a program for studying interactions between molecules, version 4.6; University of Cambridge, 2006.

(97) Smith, G.; Lynch, D. E.; Byriel, K. A.; Kennard, C. H. L.; Colin, H. L. *J. Chem. Crystallogr.* **1997**, *27*, 307–317.

(98) Kuhn, F. E.; Groarke, M.; Bencze, E.; Herdtweck, E.; Prazeres, A.; Santos, A. M.; Calhorda, M. J.; Romao, C. C.; Goncalves, I. S.; Lopes, A. D.; Pillinger, M. *Chem. Eur. J.* **2002**, *8*, 2370–2383.

(99) Jackisch, M. A.; Butler, L. G.; Fronczek, F. R. Private communication to the Cambridge Structural Database, 2006.

CT8004326